DOCUMENT RESUME

ED 246 637 EC 170 029

AUTHOR Kratochwill, Thomas R.; Cancelli, Anthony A.
TITLE Nonbiased Assessment in Psychology and Education. Volumes I and II. [Final Report].
INSTITUTION Arizona Univ., Tucson. Coll. of Education.
SPONS AGENCY Special Education Programs (ED/OSERS), Washington, DC.
PUB DATE Nov 82
GRANT G008100160
NOTE 603p.
PUB TYPE Information Analyses (070) -- Reports - Descriptive (141)

EDRS PRICE MF03/PC25 Plus Postage.
DESCRIPTORS *Disabilities; Elementary Secondary Education; Evaluation Methods; History; Minority Groups; *Psychology; *Special Education; *Student Evaluation; *Test Bias; Testing; *Testing Problems; Test Interpretation; Test Use; Test Validity

ABSTRACT

The document presents findings from a comprehensive review of the literature on the topic of nonbiased assessment. An introductory chapter describes the review's conceptual framework. Chapters 2 through 9 present analyses of the following major aspects of the topic (sample subtopics in parentheses): historical perspectives (ancient influences, nineteenth century developments, the emergence of differential psychology); conceptual models of human functioning (seven models of human behavior that influence contemporary assessment practices); technical test bias (implications of validation theory, external and internal construct bias); situational bias in psychological assessment (test-wiseness, examiner sex and race, motivational factors); outcome bias (prediction of specific outcomes, selection versus intervention, a variety of selection models); proposed alternatives to traditional assessment (culture-reduced testing, renorming, adaptive behavior assessment, Piagetian assessment procedures, learning potential assessment, clinical neuropsychological assessment, behavioral assessment strategies); ethical and legal considerations related to nonbiased assessment of children with learning and behavior problems (moral principles, invasion of privacy, consent, access to records, issues in intervention); and the influence of professional organizations on assessment bias (positions of various professional groups regarding testing/assessment practices). A final chapter summarizes the state of the art; considers implications for decisonmaking in special education, and offers guidelines for practice. (CL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
● Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED246637

ARIZONA CENTER FOR EDUCATIONAL EVALUATION AND MEASUREMENT

Nonbiased Assessment in

Psychology and Education

Vol. I

November, 1982

THE UNIVERSITY OF ARIZONA
COLLEGE OF EDUCATION
TUCSON, ARIZONA [illegible]

EC170029

This project has been funded at least in part with Federal funds from the Department of Education, Office of Special Education, under Grant Number G008100160, Grant Authority CFDA: 84.023H. The contents of this publication do not necessarily reflect the views or policies of the Department of Education, Office of Special Education, nor does mention of trade names, commerical products, or organizations imply endorsement by the U.S. Government.

Nonbiased Assessment in Psychology and Education

Vol. I

November, 1982

[Final Report]

Thomas R. Kratochwill
The University of Arizona

Anthony A. Cancelli
Fordham University

Table of Contents

		Page
Chapter 1 -	Conceptual Framework	1
Chapter 2 -	Historical Perspectives on Assessment Bias	24
Chapter 3 -	Conceptual Models of Human Functioning: Implications for Assessment Bias	56
Chapter 4 -	Technical Test Bias	125
Chapter 5 -	Situational Bias in Psychological Assessment	170
Chapter 6 -	Outcome Bias	218
Chapter 7 -	Proposed Alternatives to Traditional Assessment	269
Chapter 8 -	Ethical and Legal Considerations	384
Chapter 9 -	The Influence of Professional Organizations	447
Chapter 10 -	Current Status	472
References		489
Footnotes		593

Chapter 1

Conceptual Framework

The use of psychological and educational tests has increased rapidly over the past few decades. This proliferation is in part documented by the growth of the Buros' Mental Measurement Yearbook from 400 pages in 1938 to over 2,000 pages in 1978 (Haney, 1981). Many reasons can and have been offered for this growth. It has been suggested that the growth of psychological testing may reflect an aspect of our mass society, specifically, a need for our institutions to deal with large numbers of individuals. With regard to the use of tests in education, Garcia (1981) concludes that proponents of mental measurement believe that "[m]easuring human abilities by standardized tests would presumably increase educational productivity and sort the various grades of humans for their roles in the industrial society" (p. 1172). It has also been suggested that the reason may, in part, be explained by the love affair that America has always had with technology and psychological testing is but one expression of that special devotion (Boorstein, 1974, cited in Haney, 1981). As Haney (1981) concludes, however, all the reasons offered for why psychological and educational testing has grown so rapidly, none attribute it to the increased ability of psychological science to better measure mental processes. Given the importance of social utility as an explanation for the growth of psychological testing rather than increased quality of the devices themselves, it is no wonder why psychology often finds itself defensively engaged in research, after the fact, to demonstrate the utility of its tests in the face of public concern. One such concern that has been heard most vociferously in recent years is the perceived biased nature of these tests when used in making decisions about individuals whose backgrounds are different from those raised in mainstream Anglo or dominant culture.

In sorting through the voluminous literature to identify the reasons for this concern, two issues stand out most prominently. First, it is suggested that our society is highly sensitized to any institutionalized practices that result in a reduction of freedom of choice (Haney, 1981). When tests are employed to make decisions that may hold back individuals or groups of individuals from sharing in what is often referred to as "the American dream", potential biases in that process that may result in an unwarranted denial of freedom are closely scrutinized. Cole (1981) points out that the concerns the public have towards psychological testing ultimately focus on the social policy decisions that are made with the aid of these tests. The use of psychological testing in the schools for the classification and placement of children in classes for the mentally handicapped serves as a case in point. Classes for the mentally handicapped have come to be known as classes having little academic emphasis, poor facilities, and inadequately trained teachers (McMillan, 1977). Given such a perspective, the use of psychological testing for placement in these often called "dead-end" educational tracks has received much scrutiny. This sensitivity combined with the Riverside epidemiological studies in the early seventies (Mercer, 1970, 1973) highlighting the disproportional representation of culturally different children in these classes resulted in demonstrative outcries with the debate carrying over into the courts.

A second reason for the public's concern for bias in testing lies in the implications that are inherent in the measured differences between culturally different children and those from the dominant Anglo culture (Reschly, 1981). Psychological tests are primarily designed to measure constructs. Consequently, one implication inherent in the design of a test purported to be valid is that variation in performance connotes differences in the measured construct. Such an implication would lead one to the inevitable conclusion that culturally different children, as a group, are less capable than Anglo children. When dealing with a construct such as intelligence that has long been viewed as a characteristic heavily influenced by genetic endowment, one can appreciate the reason for the concerns of the public regarding the potential bias in testing. The stigma attached to labeling a disproportionate number of culturally different individuals retarded and the perceived insulting nature of a premature conclusion that one group of people is less intelligent than another is not only cause for concern but to some, reprehensible.

In response to concerns for potential bias in testing, there are those who conclude that mean differences across groups is enough to substantiate charges of bias (Alley & Foster, 1978; Chinn, 1979; Hillard, 1979; Jackson, 1975; Mercer, 1976, Williams, 1974). For example, Alley and Foster (1978) conclude that for tests to be nonbiased, they need to yield equivalent distributions of scores across groups. Others have studied the question of bias by examining both tests and the assessment process in an effort to determine if measured group differences are "real" differences. Some have focused on bias in the technical validity sense, some have looked at bias as a function of situational factors inherent in testing settings, while others have focused on potential bias in the assessment process within which testing is often an integral part.

Still others have addressed concerns for bias by devoting their efforts to proposing and examining alternative methods to traditional testing practices that purport to be either inherently nonbiased (e.g., criterion-referenced testing) or improvements to present practice (e.g., renorming). In addition to research efforts, a by-product of the public's concern has been the evolution of a literature related to the judicial and legislative impact of possible bias in psychological assessment. Official positions have also been adopted by several organizations whose members are involved in psychological and educational assessment.

As described above, the response to charges of biased assessment has been a frenzy of study of the issues in a variety of disparate areas. Given the inherent unwieldy nature of the literature, periodic reviews primarily designed to allow for reflection and planning for the future become increasingly important. It is the major purpose of the report to do just that. Specifically, the purposes of the present review are fourfold. First, it is the purpose of this review to be comprehensive in scope. To that end, all the various and disparate ways that the issue of nonbiased assessment has been addressed are included. Second, an attempt is made to provide a conceptual framework for organizing the mass of information presently available on the topic. Third, a critique is offered of the writings and research in each of the areas within the framework presented. Finally, an evaluation of each of the areas within the framework will be offered to provide an opinion as to the future contribution that each has yet to make when examined against the evolving trends in the overall area of nonbiased assessment.

Conceptual Framework

When conducting a review of any body of literature, a primary goal is to develop a framework for conceptually organizing the mass of information potentially available for inclusion. It was an assumption of the present effort that the literature should, as much as possible, dictate the scope of the framework rather than the preconceived biases of the authors. Consequently, a tentative conceptual framework was postulated at the outset of the review that was, by design, continually revised during an examination of the literature.

In an effort to be as comprehensive as possible, the authors allowed their initial search to range to any literature that purported to be related to the topic of nonbiased assessment. Searches were conducted in various disciplines including education, law, sociology, psychology, and medicine with the only restriction being that the focus of the literature had to be on measures of mental and/or psychological processes and/or related behavior as they apply to decisions of selection and/or intervention.

The product of this effort is a conceptual framework that includes eight major areas, each reviewed in the remaining chapters of this report. These major areas include: (1) historical perspectives, (2) conceptual models, (3) technical test bias, (4) situational bias, (5) outcome bias, (6) proposed alternatives to traditional practice, (7) judicial and legislative influences, and (8) professional association influences. Discussed below is each of the eight major areas including a brief discription of the literature covered and the rationale for its inclusion.

Historical Perspectives

The first of these major areas, historical perspectives, reviews the evolution of psychological and educational assessment and reports on historical references to biases throughout its developing history. In order to gain a full appreciation of the issues involved in present day concerns, it is necessary to acquire an understanding of the development of psychological assessment and its relationship to the issues from which it was spawned.

Such a trek into antiquity highlights the fact that many of our present day concerns with regard to bias have their origin in past assessment practices. In addition, this excursion provides us with a fuller appreciation of the various ways cultures have conceptualized human behavior and how contemporary assessment practice has been influenced by such conceptualizations.

Conceptual Models

The second major area making up the framework for the present review involves a discussion of the conceptual models that are presently proposed for understanding human functioning. The medical, intrapsychic disease, psychoeducational process, behavioral, sociological deviance, ecological and pluralistic models are reviewed. Different models for conceptualizing human functioning each have their own assumptions regarding the "why" of behavior. Each model dictates different assessment approaches, each with different implications for bias. Consequently each of these models are described and their implications for bias in their respective assessment practices are discussed.

Empirical Studies in Bias

One of the more difficult aspects of our task was to come to grips with the various ways in which bias has been defined and consequently studied in the empirical literature. Most issues dealing with nonbiased assessment are emotionally charged and full objectivity is like the proverbial end of the rainbow - never reached. Yet, the application of scientific method to the study of bias in assessment has been both plentiful and fruitful. This gives testimony to the power of science in mediating disputes and sorting through biases even when those biases are held by those charged with applying its method. As Cole (1981) points out, researchers in this area usually fall into one of two camps; either the defenders or reformers of tests. Yet, as we will see in these areas of the review, biases notwithstanding, some definite progress has been made.

The third and fourth areas of the review reflect our conceptual organization of empirical literature that has attempted to address the question as to whether differences among groups in their performance on tests is a function of bias or represent "real" differences. Employing traditional validation theory as a basis for determining bias, these studies attempt to address the issue of whether or not tests are measuring the same construct across groups.

The fifth area also reports on an empirically-based literature that focuses on potential bias in the outcomes of the entire assessment process that either may or may not include the use of tests. These studies can be viewed as employing an expanded version of validation theory that includes the study of whether or not tests are equally valid across groups when used to predict desired outcomes. These three areas of the review are identified as technical test bias, situational bias, and outcome bias. A brief description of each follows.

Technical Test Bias. By far the most organized search for bias in assessment has come out of the literature on technical test bias and this makes up one of the major areas of this review. Technical test bias is defined as bias in a purely statistical sense. When speaking of testing, bias refers to "systematic errors in the predictive validity and construct validity of test scores of individuals that are associated with the individual's group membership" (Jensen, 1980, p. 375). Thus, those who choose to examine bias from a technical psychometric point of view have argued that bias in tests can be evaluated by using validity criteria in the examination of test responses across groups. If a test is biased then the validity of the test will be different for different groups. Making use of the notion that all forms of validity are aspects of construct validity (Cronbach, 1980; Messick, 1975), the study of technical test bias can be classified as two types: External construct bias and internal construct bias.

External Construct Bias. External construct bias can be viewed as bias determined by a criterion external to the test. It employs criteria to those used in predictive validity studies and asks the question, "Does the test relate to external criteria equally well across groups as would be predicted given the construct it purports to measure. It is argued that the fact that members from different groups perform differently on a test does not indicate bias. If one suspects bias as a consequence of differential performance among groups, then one can test the hypothesis to see if the test predicts some criterion differently for different groups. If the tests predictes differently, the test can be considered externally biased.

Internal Construct Bias. While external construct bias focuses exclusively on an external criteria to determine bias, internal construct bias focuses on the internal structure of the test to determine if the test is measuring the same thing for all, regardless of group mean differences. The fact that a test predicts equally well for all provides only partial verification that the construct the test purports to measure is doing so in an unbiased manner. In internal construct bias, methods usually used to support the construct and content validity of tests are employed to determine if such evidence is different for different groups. Factor structure bias, distractor bias and item bias have all been studied and are reviewed in this section. When group differences are found, the test can be said to be internally biased.

Situational Bias. Another area that has received a substantial amount of attention over the years is often referred to as situational bias. Situational bias, also referred to as atmosphere bias, involves the study of those influences in the test situation that may interact with group differences to produce systematic bias in performance across groups. Jensen (1980) identifies six sources of potential situational bias that have been studied and are included in this fourth major area of the review. These include (1) the effects of prior practice or coaching; (2) interpersonal factors involving the attitude, expectancy and dialect of the examiner and the manner in which the examinee is motivated to perform; (3) individual versus group administration and how general classroom morale and discipline may influence performance; (4) timed versus untimed tests; (5) the interactions with race and sex of examiner and examinee; and (6) the potential biasing influence of the halo effect and its influence on scoring test performance.

This area, like technical test bias, can be viewed as a study of the validity of tests for use with culturally different populations. Although independent of the test itself, situational factors have the potential of impacting on test scores and consequently influencing the validity of the construct. This literature makes the critical distinction between performance and capability and asks the question whether or not the performance on tests of individuals from differing cultures are accurate reflections of their capabilities (Henderson & Valencia, in press). From a social learning perspective, only when motivational conditions are optimal will the performance capability of an individual equal their actual performance. Consequently, if variations in the performance of individuals from different cultures can be manipulated by situational factors then the measure may be said to be biased. In essence, this is a test of the construct validity of the test. That is, is the test measuring the same construct equally well for all individuals?

Such information is different from that gained from examining internal construct bias since those methods can only provide evidence from which one can infer if the same construct is being measured - not how well it is being measured. Evidence of whether or not one group's performance is influenced by situational variables would, likewise, not be necessarily evidenced in an examination of external construct bias. If a situational factor (e.g., achievement motivation) influences a criterion measure (e.g., academic achievement) to the same degree that it influences a construct measure (e.g., intelligence), and if the situational factor differs among groups, then one would expect the construct measure to predict the criterion measure equally for all groups even though the construct may not be measured equally well for all groups.

Outcome Bias. As implied in our previous discussion, when the issue of nonbiased assessment has been addressed in the past, attention has usually turned to the study of tests and their validity as defined within the scope of content, construct and criterion-related validity. Yet, as defined, a technically valid test provides us limited information on its usefulness. Indeed, validity in a traditional sense tells us only how well a construct is being measured, not how useful the measure is in making decisions.

To this point both Cronback (1980) and Messick (1975) have emphasized that the different types of validity generally offered to substantiate the integrity of tests are all aspects of construct validity. The use of external criteria to establish a construct's relation to other variables is to provide further information regarding whether or not the measure is "acting" the way it is hypothesized. It is not intended to provide specific information regarding the value of its use for making a particular type of decision. This point is exemplified by the rather general nature of the criterion measures usually chosen to establish predictive validity. For example, in providing evidence for the criterion-relatedness of a measure of intelligence, a general measure of academic achievement is often used. Such a measure only implies the construct's usefulness in educational decision making. Whether or not the construct measure is useful in any one particular circumstance has to be determined by the success of the outcome predicted by its use. The validity regarding the usefulness of a test, then, involves additional information than that provided in predictive validity studies. In addition to the traditional psychometric properties of the measure, one would need to know how well the measure predicts the criterion of concern in the setting and for the individual for whom you are making the decision. If intervention planning is the purpose of the use of the measure, then one also needs to know the extent to which one can predict success of an intervention designed from the use of the test.

In discussing this point, Cole (1981) suggests that the inability of technical test validity to provide information regarding all types of interpretations that can be drawn from a test provides evidence of the limitations of validation theory. Cronback (1971) writes:

> narrowly considered, validation is the process of examining the accuracy of a specific prediction or inference made from a test score . . . More broadly, validation examines the soundness of all the interpretations of a test (p. 443, cited in Cole, 1981).

When we broaden our focus of attention in a search for bias to include the study of outcomes, it becomes readily apparent that the search needs to encompass more than the study of test bias even when conceived in its broadest sense. The information provided by tests is only one aspect bearing on the validity of decision-making and its subsequent outcomes. Additional data brought to bear on the decision-making process may include other data on psychological functioning that has no established reliability and validity (e.g. subjective judgments of a teacher regarding the intellectual functioning of the child) as well as philosophic, legal, social, and economic factors. All have their impact on intended outcomes, and all have the potential of being biased.

With respect to the latter, there are those who point out (e.g., Messick, 1975) that while there are numerous data in the decision-making process for which technical validity can be offered, there are other data related to the values of the decision makers and those responsible for the decisions made that cannot be validated in a psychometric sense. The influence brought to bear on the decision by these factors can only be judged by the potential consequences its use will have in terms of social value. Cole (1981) points out that while an intelligence test may be valid for helping in the diagnosis of the mentally retarded, "validity theory does not say whether the use of the test, or the whole system in which the test use is embedded, will produce a social good or a social evil" (p. 1068).

To further illustrate the complex of factors that may impact on an intended outcome, let us examine a typical process used to decide on the diagnosis and placement of a child suspected of being "mentally retarded". In order to engage in such decision-making, a wealth of data are usually generated. Formal test data will typically include information on the child's learning potential, adaptive behavior, and academic functioning. Other test data regarding the child's perceptual and/or social-emotional functioning may also be collected. Yet, when the ultimate decision is made regarding the classification and placement of the child, the test data becomes only one source of data used to make decisions. Nontest data may include the child's history of school performance, attendance history, attempts to remediate the problems in the mainstream class, type and quality of alternative placements available, subjective impressions of the team members regarding the child's intellectual functioning, the parent's support for diagnosis and placement, whether or not the placement involves changing schools, available transportation, number of children previously placed in such a class, whether or not such a diagnosis and placement will disrupt the proportional representation of minorities in special education, among others.

In addition to these (and many more) data that are typically included in a decision-making process are the inferences that one makes about the data themselves. Because an IQ test has been validated by demonstrating its technical validity including its relationship to academic achievement, does not necessarily make it applicable for helping any one child in any one setting. In this respect, validity questions include whether or not the external criteria used to validate the test really speak specifically to the decision made about that child in that situation. And what about the outcomes of that placement? Does the use of the assessment data, both test and nontest, lead to a decision that will help that child? If the stated purpose for collecting the data is to improve the learning of that child then one can effectively argue that the utility of the data in bringing about this outcome should be an important aspect of the validity of the data. As a consequence of the above concerns, there are those who advocate that our definition of validity be expanded to include the validity in predicting desired outcomes. From this perspective our study of bias needs to encompass all aspects of the process lending information for making the decision.

Those who advocate such a position have expanded our arena of empirical efforts in two ways. First, they have required that we clearly distinguish between what we actually know about the measures we're using (i.e., psychometric properties) and what we are inferring in any given decision-making circumstance. Second, they have broadened our study of various types of data, including nontest data, that are used in the decision-making process. This has focused our attention on how all sources of data, and the interactions among the data, influence decisions and subsequent outcomes.

Those who share this perspective usually hold to a more decision-theoretic model of assessment as opposed to a classical test-based model (Cronback, 1971). In the latter approach to assessment, emphasis is placed on the accuracy of measurement. It endorses the use of the best instrument available for measuring a construct regardless of the decision one needs to make with the data. From this point of view, if one were interested in the measurement of intelligence, for example, it wouldn't matter the purpose, the measurement of choice would be that instrument that performs the task most reliably and validly; that is, the instrument that boasts the best psychometric properties. From the perspective of this model, if a classification and subsequent placement in a special education class proved to be unsuccessful for a child, the problem would not necessarily focus on the test data. A test can be valid regardless of ineffectual outcomes. Those who hold this position usually advocate restricting our definition of bias to technical test bias. The effectiveness of an intended outcome as it includes philosophic, legal and other such considerations or the inappropriate use of test data or other data is an issue of "fairness" and "misuse", respectively, not bias.

As stated above, those who argue for a more encompassing definition of bias, tend to hold more of decision-theoretic model of assessment. From this perspective, the focus of any assessment, by its nature, is on the outcomes of the entire process. Information derived from both test and nontest data, as well as social value considerations, are all an integral part of the assessment process that cannot be divorced from the utility of the outcomes. There is no such thing as the perfect test for measuring anything. The only way one can decide on the appropriateness of a test is to view it as part of a comprehensive strategy for assessing individuals for making specific decisions. The validity of a measure must, therefore, be judged on the effectiveness of the outcomes of any decision that employ the measure in this process. Consequently, the same test or other pieces of data may be valid for making some decisions while invalid for making others.

Two different types of outcomes and consequently two different types of validities can be described. The first related to the selection of individuals, the second, intervention with individuals. Likewise, two different types of bias can be identified: (1) bias in selection and (2) bias in intervention.

Bias in Selection. The major difference between bias in selection and bias in intervention lies in the purposes of the assessments. When using tests for selection, the purpose is to identify a test or tests that will allow one to choose among those who take the test(s). So, for example, when using a test for selection in hiring or admissions, ones purpose is to choose among prospective applicants those one wants to hire or admit and those one doesn't. Since the purpose of the testing is to hire or admit those who will succeed and not hire or deny admittance to those who will not succeed, the focus is on the utility of the test in increasing the probability of making the correct choice. Bias in selection, then, relates to whether or not the decision-making is biased in selecting among all who apply, regardless of group membership.

Bias in Intervention. Tests used for decisions involving intervention, on the other hand, have an entirely different purpose from those used in selection. The ultimate purpose of this type assessment is to provide help to the individual taking the tests. More often than not, formal help is given through diagnosis and placement. For example, when children are assessed to determine special education eligibility, any subsequent diagnosis and placement decisions are ultimately made with the desired outcome of helping the children. While this may appear on the surface to be a decision involving selection (i.e., you select those who are eligible and deny those who are not), it is not. The difference is that in the intervention process, after the assessment is conducted, something active occurs (i.e., some form of treatment) that is a direct consequence of the assessment process. In selection, the assessment process stops with the decision. Our concern with testing for intervention, then, involves how well the test predicted the success of an intervention that is planned with the use of the test. As can be seen, testing for intervention involves a whole new set of inferences that must be drawn from assessment data. Bias in intervention therefore, relates to whether or not the decision-making is fair in helping all those who are assessed, regardless of group membership.

Proposed Alternatives to Traditional Practice

In response to alleged bias in tests and/or in the assessment process, a variety of procedures have been proposed as alternatives to traditional practice. These alternatives include procedural approaches that are usually represented by both test and nontest-based methods. Some are designed specifically to address the issues raised in the assessment bias literature. Others provide alternatives that have not gained popularity in traditional test practice but have been identified as yielding results that are less biased than those procedures more commonly employed in psychoeducational or psychological testing. Some are well founded procedures boosting good psychometric properties while others, at best, can be considered experimental. In addition, the various alternatives that have been proposed differ radically in the type of data they provide and, therefore, the purposes for which they can be used. The alternatives reviewed herein, run the gamut from procedures that look and act much like those typically employed in traditional practices to those that are radical departures.

These include: (1) culture-reduced testing, (2) renorming, (3) adaptive behavior measures, (4) Piagetian strategies, (5) learning potential assessment, (6) diagnostic clinical teaching, (7) child development observation, (8) psychoneurological assessment, and (9) behavioral assessment. Included in the discussion of behavioral assessment is criterion-referenced testing.

Culture-Reduced Tests. These tests are sometimes purported to contain content that is either free of culture or fair to individuals regardless of the culture in which they are a member. The aim of culture-reduced tests is to include content influenced only by environmental circumstances that are common across cultures. Cattell's Culture-Fair Intelligence Test is an example of those included in this type. Also included under the heading of culture-reduced tests are nonverbal tests. Nonverbal tests are those tests that purport to be language-reduced. Some eliminate requirements for verbal responding while others eliminate both verbal instructions and verbal responding as prerequisities to performance. These tests are reported to be less biased with multilingual and some physically handicapped individuals. Several of these tests are now appearing on the market in response to charges of language bias in testing while others have been available for years for use primarily with handicapped populations. The Nonverbal Test of Cognitive Skills is an example of a type nonverbal test included in this category.

Renorming. Renorming involves taking an already established test and providing new norms that are more characteristic of the population of individuals being tested than the national representative samples that are most often used to originally norm the test. This alternative is most prominently embodied in the SOMPA.

Adaptive Behavior Measures. With a reconceptualization of the meaning of adaptive behavior as evidenced in the 1977 revision of its definition by the American Association of Mental Deficiency, measures of adaptive behavior have gained a resurgence of popularity in recent years. Within this new conceptualization, adaptive behavior during childhood includes the child's adaptation to the community as well as the school (Reschly, 1982). This reconceptualization was in part, if not entirely, the result of potential bias in the diagnosis of mental retardation when assessment was conducted, as previously had been done, by an examination of adaption to the school culture only. The measurement of adaptive behavior, by law, is now a necessary component of any diagnosis of mental retardation. An example of a new measure of adaptive behavior designed specifically to address the problem of bias in the diagnosis of mental retardation is the Adaptive Behavior in Children (ABIC) scales which is part of the SOMPA.

Piagetian Tests. Many of the procedures used to measure constructs employed in Piaget's theory of intellectual development are less than more traditional tests of intellectual functioning. The unique feature about Piagetian tests that make them candidates for alternative nonbiased procedures is in the nature of the constructs they purport to measure. Reported by those involved in this area of research, the constructs are purported to be universal and invariant. Some evidence has been reported regarding the similarity of cognitive development, as defined by these measures, of children from diverse cultural backgrounds (c.f. De Avila, & Harassy, 1975).

Learning Potential Assessment. Learning potential assessment procedures involve a test-teach-test model of assessment that differs dramatically from traditional measures that sample behavior at one point in time. Given the fact that a teaching component is provided in the assessment, it is reported by some to be less influenced by the cultural background of the child. Since this type assessment has been developed most extensively by Fuerstein (1978) as a component of an intervention program, its procedures are less standardized than normally found in traditional tests.

Diagnostic-Clinical Teaching. Diagnostic-clinical teaching is an assessment procedure that involves the actual teaching of curriculum-related materials under conditions that maximize learning. These conditions can include a variety of manipulations such as varying reinforcement and feedback conditions. Kratochwill et al. (1980) report its relevance to nonbiased assessment as a consequence of its focus on (1) tasks that nearly all children experience in the school curriculum and (2) its relationship to the interventions that are planned from it.

Child Development Observation (CDO). Most closely associated with Ozer and his associates (Ozer, 1966, 1968, 1978), CDO is designed to simulate the process of learning on protocols that sample conditions under which a given child's learning problem may be solved. It is a nontraditional form of assessment and its non-normative approach makes it an eligible candidate as a nonbiased alternative to traditional practice.

Clinical Neuropsychological Assessment. Clinical neuropsychological assessment is concerned with the assessment of brain-behavior relations. As such, it can be conceptualized as a set of procedures best interpreted within the framework of the medical model. According to Mercer (1979), measures interpreted within this framework bear the potential of being nonbiased since the norms developed for such measures are minimally influenced by variations in culture. The procedures themselves depend on standardized behavioral observations used in conjunction with normative psychological assessment devices.

Behavioral Assessment. Most commonly associated with behavior therapy approaches, behavioral assessment has been identified with nonbiased assessment since it involves a set of procedures that sample behaviors that are most often referenced to an absolute standard of performance. The sample is usually taken in the natural environment and the desired standard of performance established with either a person responsible for the individual's behavior or the individual himself/herself.

Criterion-Referenced Tests. While not originally designed specifically as nonbiased measures, the assumptions underlying the development of criterion-referenced tests make them candidates for such use. This class of measures, unlike traditional norm-referenced tests, do not depend on comparing children in the assessment of abilities and skill-level achievement. Instead, criterion-referenced tests measure the extent to which a child has mastered an absolute preestablished standard of performance. These tests are sometimes referred to as domain-referenced tests.

Judicial and Legislative Influences

One by-product of the public's concerns over perceived bias in psychological and educational assessment has been the involvement of the judicial and legislative systems in attempting to resolve the issue. The present review focuses primarily on those legislative and judicial actions that have impacted on testing as it relates to bias in educational practice. However, since in the area of judicial actions the influence of rulings in one area of the application of testing are felt in all areas, the discussion will also extend to those court cases that have had an indirect, yet significant, impact on assessment bias in education. This is especially true of those rulings on tests in the area of employment.

The impact of legislative and judicial actions on psychological testing in education should not be underestimated. Since the mid-1960's, a wealth of litigation and legislation has evolved that have affected the administration, interpretation and use of psychological tests (Bersoff, 1981). Legislative actions such as the Civil Rights Act of 1964 and P.L. 94-142 are two of the more prominent laws that are presently reviewed.

In the area of judicial action, the courts, who have traditionally attempted to maintain a "hands-off" posture with respect to issues involving school policy, have recently jumped into the arena "with both feet". Hearing cases on both statutory and constitutional grounds, the courts have steadily increased their involvement in the fair use of psychological testing and will apparently continue to do so (Bersoff, 1981). For the purposes of this review, major attention has been focused on the Larry P. v. Riles (1979) and PASE v. Hannon (1980) cases. These cases most directly impact on the use of intelligence tests for the diagnosis and placement of children in "educable mentally retarded" classes.

Professional Association Influences

The eighth and last major area within the conceptual framework that will be reviewed, involves the influence of professional associations on the biased assessment practices of its members. The impact of professional associations is usually felt through training programs, public statements, published guidelines and their impact on the certification and licensure of those who qualify to administer tests or provide professional services. In this area of the review some of the professional groups that have set forth standards for assessment practices for its members will be examined.

Structure of the Report

Our review of assessment bias is composed of 10 chapters. Chapter 1, the present chapter, was designed as an introduction to the report. Its purpose was to detail the conceptual framework that evolved as a consequence of our review and that has provided the basic structure for this report. Chapters 2 through 9, inclusive, contain a discussion of the major areas reviewed. Chapter Two reports on the historical perspectives to bias in assessment while the conceptual models of human functioning are reviewed in Chapter Three. Chapter Four, Five, and Six report on empirical studies in assessment bias in the areas of technical test bias, situational bias, and outcome bias, respectively. The various alternatives to traditional test practice are reviewed in Chapter Seven. Chapter Eight reviews the legislative and judicial influences on bias in testing and the influence of professional organizations is reported in Chapter Nine. In Chapter Ten, a synthesis of all the major areas is attempted to provide the reader various perspectives on where we have come in our understanding of bias in assessment and, more importantly, where we still have to develop new areas of research and practice.

Chapter 2

Historical Perspectives on Assessment Bias

Since the beginning of assessment efforts, individuals have been concerned with how fair the actual procedure or technique was for those participating in it. In this chapter we trace the development of assessment over recorded history up to the present. Although our overview is quite focused (see several sources for a more detailed general review: DuBois, 1970; Doyle, 1974; Linden & Linden, 1968; McReynolds, 1975), we provide a perspective on comtemporary bias in the assessment process.

An examination of the historical factors in assessment is important for several reasons. First, it is important to understand that many of the contemporary issues in assessment bias have their origin in past assessment practices. Second, it is important to realize that many contemporary issues are related to social or even political concerns that have their origin in the past. Third, the past has sometimes provided or even imposed a structure on assessment practices. It is important to understand this structure in order to identify contemporary models of assessment practice. Finally, it is important to focus on historical factors to introduce a variety of scholarly perspectives into the discussion of the issues surrounding bias in assessment.

Ancient Influeces

One of the most extensive and scholarly discussions of the historical antecedents of assessment in general and personality assessment specifically, has been presented by McReynolds (1974). Most historical treatments of the assessment literature typically begin with a discussion of the work of Galton in England and Catell in the United States [i.e., many books on assessment begin with this period (e.g., Sundberg, 1977)] and historical tables reflect this perspective. However, assessment has a much richer history, attesting to the assumption that many features of contemporary assessment actually date back to the beginnings of recorded history. McReynolds (1974) traced the historical antecedents of the current practices in assessment beginning with antiquity and extending to the latter part of the last century. Four phases are reviewed, namely, antiquity, the medieval period and the Renaissance, the Age of Reason, and the period from Thomesius to Galton.

Antiquity

An examination of early assessment practices shows that there was a close interplay between the methods employed and the cultural views held during that particular time. This is not unlike the contemporary views in the United States that led to the development of PL-94-142 with its emphasis on fair assessment pratices for handicapped children. It is possible that the first personality assessment procedure was based on astrology, and that the first psychological "test" was the horoscope. Although astrology can be regarded as invalid on scientific grounds (and possibly a biased assessment procedure), it did contribute to (a) the view that individual personalities represent the focus of assessment, (b) the psychological make-up of the individual is predetermined, and (c) the development of taxonomical categories.

Another early assessment strategy involved physiognomy, the interpretation of an individual's character from body physique. Physiognomics, also a very limited assessment procedure, assumed a relatively fixed conception of personality, but shares some methodological features with contemporary naturalistic observation, as represented in behavior modification procedures (see discussion in Chapter 7). McReynolds (1974) noted that the longest continued assessment technique with some claim to rationality and one that remains with us today is physiognomy. Thus, recent work such as that by Mahl (1956) and Gleser, Gottschalk, and Springer (1961) on speech patterns; by Hall (1959), Eibl-Eibesfeldt (1971) and Haas (1972) on ethology of movements; of Izard (1971) and Ekman and associates (Ekman 1973; Ekman, Friesen, & Ellsworth, 1972) on emotions and facial expressions; and Hess and associates (Hess & Polt, 1960; Hess, Seltzer, & Schlien, 1965) on the relation of pupil size to affect, can be related to earlier physiognomic conceptions (cf. McReynolds, 1974).

Developments in assessment during early times were not always limited to the area of "personality assesment". For example, Civil Service examinations were used in ancient China for selection purposes. DuBois (1966) notes:

> The earliest development seems to have been a rudimentary form of proficiency testing. About the year 2200 B.C. the emperor of China is said to have examined his officials every third year...

> A thousand years later in 1115 B.C., at the beginning of the Chan dynasty, formal examining procedures were established. Here the record is clear. Job sample tests were used requiring proficiency in the five basic arts: music, archery, horsemanship, writing, and arithmetic....Knowledge of a sixth act was also required - skill in the rites and ceremonies of public and social life (pp. 30-31).

Medieval Period and the Renaissance

McReynolds (1974) notes that during this period, the acceptance of humoral psychology and physiognomic strategies of evaluating people was widespread. Generally, this period supported the recognition of the individual and so we again see an example of cultural influences on assessment practice.

Age of Reason

The Age of Reason covers the period from approximately the middle of the sixteenth century to the latter part of the eighteenth. A major theme of this period was the focus on individual differences as reflected in some important works on assessment--Huarte's Tryal of Wits, Wright's Passions of the Minde, and Thomesius' New Discovery. During this period, the recognition of individual differences prompted measurement so that an individual's happiness could be more fully realized.

From Thomesius to Galton

A significant contribution to assessment during this period, particularly in the nineteenth century, was phrenology. Phrenology bears similarity to physiognomy--While physiognomy emphasized assessment of external body features such as facial and other characteristics, phrenology emphasized the assessment of the external formations of the skull. However, phrenology assumed that mental functions were based on specific processes localized in certain areas of the brain and that the intensity or magnitude of these functions was indicated in the contours and external topography of the skull (McReynolds, 1974).

Four positive contributions of phrenology that have a resemblance to contemporary assessment practices or activities can be identified (McReynolds, 1974). First, there was an emphasis on individual differences. Second, the assessment paradigm emphasized the notions of assessor and subject, the systematic collections of data during a single session, and written reports which usually included qualitative profiles. Third, the phrenological movement helped advance "objectivity" through "blind assessment" and rating scales. Finally, phrenology contributed to the development of a primitive taxonomical system such as affective faculties (e.g., propensities, sentiments) and intellectual faculties (e.g., perceptive, reflective).

Implications

This brief historical overview of ancient influences points out that many contemporary assessment practices have their roots deep in our past. Noteworthy is the fact that the work of the phrenologists, (and later Quetelet's work on psychological statistics) set the stage for the emergence of Galton's contributions and the more modern era in assessment. It is interesting to speculate how some of the ancient procedures might have been perceived as biased or discriminatory. McReynolds (1974) raises as interesting point:

> We know that such techniques as chiromancy, metaposcopy, and phrenology are in principle all totally invalid, yet I suggest that in the hands of insightful and discerning practitioners they may, at least on occasion, have been more valid than we suppose, even if for different reasons than their users, much less their clients, imagined (pp. 524-525).

Nineteenth Century

During the nineteenth century significant developments were taking place in Western Europe and the United States that would shape the future of psychological and educational assessment (cf. Carroll, 1978; Laosa, 1977; Dubois, 1970). Specifically, events were occurring in France, Germany, England, and the United States that were to have a profound influence on assessment practices in psychology and education.

France

Attention to two movements occurred in France that made a significant impact on the history of testing and assessment (Maloney & Ward, 1976). One movement, pioneered by Bernheim, Liebault, Charcot, and Freud, was focused on a new view of deviant behavior. The influence of this movement was to take abnormal behavior out of the legal or moral realm with which it had been previously associated and cast it as a psychological or psychosocial problem. This prompted psychological assessment rather than moral or legal sanction, as had been common prior to this period.

Also noteworthy was the movement called "The Science of Education." Jacques Itard, a French physician, taught Victor, the "Wild Boy of Avegron" various skills. Many of the procedures used in Itard's work were similar to more contemporary behavior modification procedures which emphasize environmental stimulus and response changes during instruction. Itard's contributions also provided a background for Binet's work on measurement of intelligence.

Esquivol's (1722-1840) work, represented in his book Des Maladies Mentales was influencial in that he distinguished between "emotional disorders" and "subaverage intellect." According to his views, subaverage intelligence consisted of levels of individual performance: (a) those making cries only, (b) those using monosyllables, and (c) those using short phrases, but not elaborate speech. Thus, here we see the basis for an early classification scheme that could organize human behavior.

Germany

While some of the work in France emphasized individual differences in pathology and cognitive ability, German scientists perceived individual differences as a source of measurement error. A significant contribution to the individual differences theme is found in the "Maskelyne-Kinnebrook affair." The difference between Maskelyne (the astronomer) and Kinnebrook (the assistant) in their measurement of the timing of stellar transits was later analyzed by Bessel. Bessel concluded that different persons had different transit tracking times, and that when all astronomers were checked against one standard, individual error could be calculated--a sort of "personal equation" was developed (cf. Boring, 1950).

Another significant influence on assessment came from Wundt who set up a psychological laboratory in Leipzig to study such processes as reaction time, sensation, psychophysics, and association. This work, as well as the general work occurring on measurement was helpful to popularize the notion of measurement of differences between individuals. Some Americans who studied with Wundt were G. Stanley Hall and James McKeen Cattell. Both of these individuals were to have a large impact on future psychological assessment.

England

The work of Charles Darwin was most influential in psychological and educational assessment particularly in his theory of evolution presented in 1859 in Origin of the Species. Darwin's work emphasized that there are measurable and meaningful differences among members of each species. Galton, Darwin's half-cousin, was most influential in applying evolutionary theory to humans. In his book, Hereditary Genius (1869), he argued that "genius" had a tendency to run in families. Galton was greatly influenced by the Belgian statistician Quetelet (1770-1864) who was the first to apply the normal probability curve of Laplace and Gauss to human data. This translated into the notion of "l'homme mayen" or the notion of an "average man" (Boring, 1950). In this view, nature's mistakes were represented as deviations from the average.

Several implications of this work are noteworthy. First, Galton's system of classification represented a fundamental step toward the concept of standardized scores (Weisman, 1967). Second, in the application of Quetelet's statistics, Galton demonstrated that many human variables, both physical and psychological, were distributed normally. This is a direct precursor to the concept of a norm and application of standardization (Laosa, 1977). Third, a major influence of this work was to establish that certain variables should be subjected to quantitative measurement. Galton's work was significant in that it encouraged other efforts in the area of measurement of individual differences in mental abilities that was considerably more sophisticated than previous efforts (Cooley & Lohnes, 1976). Finally, through the application of the normal curve, individual performance or standing could be classified as deviant or even as a mistake of nature. We know that although Galton was influenced by the phrenologists, he rejected this form of assessment. He noted in 1906, "Why capable observers should have come to such strange conclusions (can) be accounted for...most easily on the supposition of unconscious bias in collecting data" (quoted in Pearson, 1930, Vol IIIb, p. 577).

United States

Early work in the United States contributed to what was called the "Mental Testing" movement. Cattell (1860-1944) was the first to use the term "mental test" and he is generally referred to as the father of mental testing (DuBois, 1970; Hunt, 1961). Cattell also introduced experimental psychology into the United States. A significant contribution to assessment was that he advocated testing in schools; he was also generally responsible for instigating mental testing in America (Boring, 1950).

In 1895 Cattell chaired the first American Psychological Association Committee on Mental and Physical Tests. Although Cattell made major changes in the nature of testing, his work was not accepted unconditionally. For example, Sharp (1899) published an article questioning the reliability of mental tests. Wissler (1901) compared the reliability of some of Cattell's psychological measures with various measuring approaches from the physical sciences and concluded that tests used in Cattell's lab showed little correlation among themselves, did not relate to academic grades, and were unreliable (cf. Maloney & Ward, 1976). Even Wundt was not supportive of Cattell's focus on mental measurements (Boring, 1950). Nevertheless, Cattell's work, as well as other work in France, promoted the development of a movement called differential psychology.

Differential Psychology

Applications in Education

Around the turn of the century, assessment was again given a new impetus through the development of differential psychology (Binet & Henri, 1895; Stern, 1900, 1914). Stern (1914) suggested that mental age be divided by chronological age to produce a "mental quotient," a procedure, with refinements, that has evolved into the IQ of today (Lasoa, 1977).

The work of Binet and his associates was quite influential, although not necessarily in the direction that Binet had envisioned or desired (cf. Sarason, 1976; Wolf, 1973). Binet initially focused his efforts on the diagnosis of mentally retarded chidren around the late 1880's. At this time he was assisted by Theodore Simon, who he later worked with in the development of the first formal measure of intellectual assessment for children (Wolf, 1973). Based on a study conducted for the Ministry of Public Instruction, he focused efforts on predicting which child would be unable to succeed in school (Resnick, 1982). Binet noted that performance on his scale had implications for classification and education. Resnick (1982) notes:

> A scale of thirty questions was developed, each of increasing difficulty. Idiots were those who could not go beyond the sixth item, and imbeciles were stymied after the twelfth. Morons were found able to deal with the first twenty-three questions. They were able to do the memory tests and arrange lines and weights in a series, but no more...the test.. was designed as an examination to remove from the mainstream of schooling, and place in newly developed special classes for the retarded, those who would be unable to follow the normal prescribed curriculum. As such, it was a test for selection, removing from normal instruction those with the lowest level of ability. Binet argued, however, that the treatment the children would receive in the special classes would be more suited to their learning needs. The testing, therefore, was to promote more effective and appropriate instruction (p. 176).

Around the turn of the century, interest in testing the abilities of children was at a high level. This was prompted, in part, by the growing population of children in schools due to natural population growth and immigration (Trow, 1966), and the fact that students began to stay in school longer (Chapman, 1979). With the growing number of children in schools, it became clear that not all children could profit from regular instruction. A Senate Committee reported in 1908 that approximately 72% of all foreign born public school students in New York [and in many other cities it was close to 50% (Tyack, 1974)] could profit from special instructions.

Several American psychologists promoted Binet's work. For example, Henry Goddard published the first revision of the Binet scale and Terman developed the Stanford-Binet. Thereafter the Binet scale was use to identify children who were regarded as "backwards" or "feebleminded". Wallin (1914) reported that in 1911 the Binet was being used in 71 of 84 cities that administered tests to identify "feebleminded" children. However, the Binet scale was also being used experimentally to screen out and turn back retarded immigrants (Knox, 1914, cited in Wigdor & Garner, 1982).

The Stanford version of the Binet-Simon Scale was originally published in 1916 by Terman and this scale was revised by Terman and Merrill in 1937 and 1960 and renormed in 1972. This translation and revision of Binet's earlier work firmly established intelligence testing in schools and clinics throughout the United States (DuBois, 1970). It is possible that work building on these develoments led directly to many of the issues surrounding bias in assessment practices in psychology and education today. Sarason (1976) notes:

> School psychology was born in the prison of a test and although the cell has been enlarged somewhat, it is still a prison. Alfred Binet would have been aghast, I think, to find that he gave impetus to a role which became technical and narrow, a role in which one came up with analyses, numbers,

> and classifications which had little or no bearing on what happened to children in the classroom. Of course, it makes a difference if, on the basis of testing, a child is put in a special class of some kind---and we certainly have a variety of types--but even here Binet would probably have asked what bearing had the child's performance on the specific educational plan which he required (p. 587).

Development of Group Testing

The assessment movement was given a major thrust through the development of group tests during World War I (WWI). Many assessment efforts during this time reflected a pattern of procedures similar to that used by Binet (Newland, 1977). Ebbinghaus demonstrated the feasibility of group tests and some American psychologists (e.g., Whipple, 1910; Otis, 1918) recognized that the Binet-Simon Scale could be adapted for group testing. However, there were important differences. Whereas the Binet-type items typically required a definite answer provided by the child, group tests usually called for recognition of a correct answer among several alternatives (Carroll, 1978).

A committee of the American Psychological Association, chaired by Robert M. Yerkes, developed the Army Alpha and Army Beta group tests. The Army Beta (a nonverbal group test) was designed so as not to discriminate against illiterates and individuals speaking foreign languages. While the impact of this development was to create a new interest and role in testing, a review of tests used (cf. Yerkes, 1921) reveals the source of many tests were increasingly used for non-military purposes (Newland, 1977).

Following the war, many psychologists who were involved in wartime testing sought employment in the civilian ranks and many became involved in the schools. Resnick (1982) notes:

> Aiding this movement was Philander P. Clarxton, U.S. Commissioner of education, who circularized school superitendents throughout the country about the reserve of trained people that could be tapped for the needs of the schools. He wrote enthusiastically about the "unusual opportunity for city schools to obtain the services of competent men..." Among the services they could render was "discovering defective children and children of superior intelligence..." (p. 183).

This movement, in part, facilitated the use of group intelligence tests in the public schools. Many of these tests were administered to identify children who could not profit from regular instruction. Although some schools had made provisions for special children (Wallin, 1914), the intelligence tests served a role to formalize the decision making process for these special services. Also, between 1919 and 1923, Terman introduced the National Intelligence Test for grades three to eight, and the Terman Group Test, for grades seven to twelve and found that the schools were most receptive (Resnick , 1982). Resnick (1982) reports that the most important use of the tests was for placement of children in homogeneous groups:

> Sixty-four percent of the reporting cities used group intelligence tests for this purpose in elementary schools, 56 percent in junior high schools, and 41 percent in high schools. Enthusiasm for the use of testing systemwide for this purpose was at a high level. In 1923, Terman's group test for grades seven to thirteen

sold more than a half-million copies (pp. 184-185).

The stage for the rapid development of ability tests was also set by such individuals as Spearman, Thorndike, and Thurstone. Spearman developed an elaborate theory of the organization of human abilities in which he concluded that all intellectual abilities have a common factor, g, and a number of specific factors, s, which relate uniquely to each presumed ability. Spearman's two-factor theory was the basis upon which tests examining specific abilities (Edwards, 1971) rather than global scores were developed (Laosa, 1977).

Thorndike viewed intelligence as comprised of a multitude of separate elements, each of which represented a specific ability. Intelligence was also perceived as having both hereditary and environmental components. Thurstone concluded that there were seven primary mental abilities (in contrast to Spreaman's s, factors) and developed the Primary Mental Abilities Test to measure each specific ability.

Intelligence tests gradually evolved into major diagnostic instruments throughout the world. Such instruments became a major diagnostic tool for identifying the retarded for psycho-educational research and service (cf. UNESCO, 1960). However, not all countries accepted their use: In the Soviet Union such tests were banned in 1936 by the Communist Party because they were considered methods which discriminated against the peasants and the working class in favor of the culturally advantaged (Sundberg, 1977; Wortis, 1960). As an alternative, diagnosis was based primarily on neuro-physiological evidence. The neurologist and psycho-physiologist, rather than clinical psychologist, were primarily engaged in diagnosing the mentally retarded (cf. Dunn & Kirk, 1963).

Work in these areas, as well as other contributions prior to and during this period, led to diverse views on the nature of intelligence and its assessment. A major contribution to the testing movement was the development of the Wechsler Scales. Wechsler developed the Wechsler Adult Intelligence Scale (WAIS) by including a group of sub-tests from WWI vintage which were found valuable in his work with adults. His criterion of "general adaptability" (cf. Wechsler, 1975) was extended downward in the development of the Wechsler Intelligence Scale for Children (WISC and WISC-R) and the Wechsler Pre-School and Primary Scale of Intelligence (WPPSI). The work of Wechsler contrasted with that of Binet. Whereas Wechsler's Scales emerged from work with adults and were later developed for use with children, Binet's emerged from work with young children and later was developed for use with older children (Newland, 1977). This has led to an important differentiation that has implications for assessment:

> The perception of tested intelligence in adults today has hampered and diluted the perception of tested learning aptitude in children. And yet, in spite of the fact that so many different measures are objectively obtained on children, such results are used in research along with those obtained otherwise on adults as though they were interchangeable (Newland, 1977, p. 6).

Newland (1977) suggests that "learning aptitude" (in the sense of school learning aptitude) is a much better criterion for "child intelligence" than the adult connotation of multi-faceted susceptibility of adaptation or potential of adults.

Political Aspects of the Assessment Movement

The testing movement has not been confined to issues bearing on the psychometric features of tests themselves. Test results and data from testing research have been used for political or even racial positions. Many European and American scientists (anthropologists, biologists, and psychologists) have held racial positions (Chase, 1977), and this has been documented specifically with testing the IQ of individuals (Block & Dworkin, 1976; Eckberg, 1979; Gould, 1978; Kamin, 1974).

Many psychologists interpreted the intelligence test data from WWI as evidence for genetic differences among races and within the Caucasian race, among different nationality groupings (e.g., Brigham, 1930). However, some of the interpretations were later retracted (e.g., Brigham, 1930). Indeed, the notion that intelligence or scholastic aptitude reflected largely the effects of native endowment in interaction with schooling was generally slow in development (cf. Carroll, 1978, e.g., Peterson, 1925).

Nevertheless, a variety of oppressive positions by "respected" individuals were presented during the history of testing, as these statements indicate:

> (W)e are incorporating the negro into our racial stock, while all of Europe is comparatively free from this taint...the steps that should be taken...must be of course be dictated by science and not by political expediency...the really important steps are those looking toward the preventions of the continued propagation

of defective strains in the present population (Brigham, 1923)

Their [Mexican and Indian children's] dullness seems to be racial, or at least inherent in the family stocks from which they come. The fact that one meets this type with such extraordinary frequency among Indians, Mexicans and negroes suggests quite forcibly that the whole question of racial differences in mental traits will have to be taken up anew...there will be discovered enormously significant racial differences which cannot be wiped out by any scheme of mental culture.

Children of this group should be segregated in special classes...they cannot master abstractions, but they can often be made efficient workers...There is no possibility at present of convincing society that they should not be allowed to reproduce...they constitute a grave problem because of their unusually prolific breeding (Terman, 1916, p. 6).

Now the fact is, that workman may have a ten year intelligence while you have a twenty. To demand for him a home as you enjoy is as absurd as it would be to insist that every laborer should receive a graduate fellowship. How can there be such a thing as social equality with this wide range of mental capacity? ...The man of intelligence has spent his money wisely, has saved until he has enough to provide for his needs in case of sickness, while the man of low intelligence, no matter how much money he would have earned, would have spent much of it foolishly....During the past year, the coal miners in certain parts of the country have earned more money than the operators and yet today when the mines shut down for a time, those people are the first to suffer. They did not save anything, although their whole life has taught them that mining is an irregular thing and that...they should save....(Goddard, 1920, p. 8)

Never should such a diagnosis [of feeblemindedness] be made on the IQ alone....We must inquire further into the subject's economic history. What is his occupation; his pay....We must learn what we can about his immediate family. What is the economic status or occupation of the parents?...When...this information has been collected...the psychologist may be of great value in getting the subject into the most suitable place in society...(Yerkes, 1923, p. 8)

Goddard reported that, based upon his examination of the "great mass of average immigrants," 83% of Jews, 80% of Hungarians, 79% of Italians, and 87% of Russians were "feebleminded" (Goddard, 1913) (in Kamin, 1975, p. 319)

That part of the law which has to do with the nonquota immigrants should be modified....All mental testing upon children of Spanish-American descent has shown that the average intelligence of this group is even lower than the average intelligence of the Portuguese and Negro children...in this study. Yet Mexicans are flowing into the country...

From Canada we are getting...the less intelligent of the

> working-class people....The increase in the number of French Canadians is alarming. Whole New England villages and towns are filled with them. The average intelligence of the French Canadian group in our data approaches the level of the average Negro intelligence.
>
> I have seen gatherings of the foreign-born in which narrow and sloping foreheads were the rule....In every face there was something wrong - lips thick, mouth coarse...chin poorly formed...sugar-loaf heads...goose-bill noses...a set of skew-molds discarded by the Creator....Immigration officials...report vast troubles in extracting the truth from certain brunette nationalities (Hirsch, 1926, p. 28).

Such positions clearly have degraded scientific attempts to deal with the nature-nurture issue. Increased controversy has surrounded such notions as intelligence being fixed and predetermined (Hunt, 1961), or being influenced by environmental or social forces. The "nature-nurture controversy" was given increased momentum in 1969 in Arthur Jensen's Harvard Educational Review article "How Much Can We Boost IQ and Scholastic Achievement" in which he discussed the relative contribution of genetic and environmental factors on IQ. Jensen (1969) indicated that (a) compensatory education for disadvantaged groups had "apparently" been a failure, (b) there was evidence to "make it a not unreasonable hypothesis that genetic factors are strongly implicated in the average Negro-White intelligence difference" (p. 82), and (c) the race differences were evident in conceptual ability (Level II) , but not in associative ability (Level I). Despite continued attacks, Jensen has defended his position (cf. Jensen, 1973a, 1973b).

Unfortunately, statements continued to support racial perspectives. Shockley (1971) noted that "Nature has color coded groups of individuals so that statistically reliable predictions of their adaptability to intellectually rewarding and effective lives can easily be made and profitable be used by the pragmatic man in the street (p. 375). Thus, although research will continue to have a bearing on issues related to test bias, the legacy from the past and present will likely influence any scientific analysis of the issues. Indeed, science occurs in a social context and it is that context that must continually be questioned (Sewell, 1981). Thus, as noted by Reynolds (1982), a greater degree of scientific skepticism may be needed for examination of the issues surrounding test bias if the errors of the past are to be avoided.

Personality Assessment Movement

Development of Traditional Tests

While tests of cognitive ability were rapidly evolving during the early part of the century, tests of "personality" were in their infancy. Although such devices as the Woodworth Personal Data Sheet were used in the military during WWI, the personality assessment movement received increased attention through the development of projective techniques such as the Rorschach and Thematic Apperception Test (TAT).

World War II (WWII), like the first war, did much to set the stage for rapid proliferation of testing practices. Indeed, psychological testing combined with the military need for assessment was one of the primary factors leading to the development of clinical psychology as an independent specialty (cf. Maloney & Ward, 1976).

During the period following WW II, testing practices developed dramatically. Most tests developed during this period were tied to an intrapsychic disease model or state-trait conceptualization of behavior (cf. Mischel, 1968). Psychoanalytic theory generally accelerated assessment procedures that would reveal unconscious processes. Assessment practices emphasized an "indirect-sign" paradigm. Assessment was indirect in that measurement of certain facets of behavior were disguised or hidden from the client (e.g., TAT). Moreover, within the context of the intrapsychic model, testing practices were said to predict certain states or traits. The clinician's task was to administer a battery of tests to a client and look for certain signs of traits or states. An example of this approach was represented in the work of Rappaport, Gill, and Shafer (1945). In their classic book the authors demonstrated how a battery of tests (e.g., TAT, Rorschach, WAIS) could be used to diagnose deviant behavior within the intrapsychic model (in this case the psychoanalytic model).

Similar to the sign approach was the "cookbook" method of assessment that reached a zenith during the mid-1950's (cf. Meehl, 1956). An example of this approach was the Minnesota Multiphasic Personality Inventory (MMPI) (Hathaway & McKinley, 1943). As these authors note, one of the presumed advantages of the cookbook approach was that "it would stress representativeness of behavioral sampling, accuracy in recording and cataloguing data from research studies, and optional weighting of relevant variables and it would permit professional time and talent to be used economically" (p. 243).

Emergence of Behavior Modification and Assessment

Behavior modification or behavior therapy and assessment affiliated with this model have made tremendous impact on psychology and education in recent years, but has been almost completely overlooked in historical accounts of bias in assessment (see, however, Kratochwill et al. 1980).

As recent historical reviews illustrate (Hersen, 1976; Kazdin, 1978) behavior therapy represents a departure from traditional models of assessment and treatment of abnormal behavior, both psychological and educational. Although the history of behavior therapy cannot be traced along a single line, contemporary *practice* is characterized by diversity of viewpoints, a broad range of heterogeneous procedures with vastly different rationales, open debates over conceptual bases, methodological requirements, and evidence of efficacy (Kazdin & Wilson, 1978). Some reports of behavioral treatment followed Watson and Rayner's (1920) work in conditioning of fear in a child, but a significant impetus to behavioral treatment is commonly traced to the publication of Wolpe's (1958) *Psychotherapy by Reciprocal Inhibition*.

Independent of Watson and Wolpe's work was research in the psychology of learning, both in Russia and the United States. Particularly important in learning research was operant conditoning which Skinner brought into focus in the late 1930s. The evolution of operant work into experimental and applied behavior analysis has had an important influence in the development of behavior therapy and assessment practices in general.

Although behavior therapy and assessment has evolved considerably over the past few years some general characteristics represents unities within the heterogeneity of contemporary practice:

1. Focus upon current rather than historical determinants of behavior;
2. Emphasis on overt behavior change as the main criterion by which treatment should be evaluated;
3. Specification of treatment in objective terms so as to make replication possible;
4. Reliance upon basic research in psychology as a source of hypotheses about treatment and specific therapy techniques; and
5. Specificity in defining, treating, and measureing the target problem in therapy (Kazdin, 1978, p. 375).

A detailed account of the history of behavior modification can be found in Kazdin (1978).

With the advent of behavior modification and its proliferation, a new assessment role also developed, particularly for clinical psychologists. Behavioral assessment emphasized repeated measurement of some target problem prior (baseline), during, and after (follow-up) the intervention. Hersen et al. (1976) note that the psychologist's expertise in theory and application of behavioral therapy techniques (e.g., classical and operant conditioning) also enabled both an assessment and treatment role to emerge in psychiatric settings. Thus, the psychologist in various settings (e.g., clinics, hospitals, schools) became involved in direct service, rather than engaged in testing and diagnosis. Behavior modification provided the impetus for these new roles.

Developments in behavioral assessment have also influenced the field of personality testing in general. In many respects assessment has acted as a barometer for the current thinking of personality theorists. For example, a barometer of change in views about assessment has been the evolution of the title of the journal specifically devoted to assessment in professional psychology (Goldfried, 1976). The journal, initially founded in 1936, was entitled *Rorschach Research Exchange*. Gradually other projective techniques came into existence in the assessment process and by 1947 the title was changed to the *Rorschach Research Exchange and Journal of Projective Techniques*. Because the Rorschach became less dominant in assessment, the name was again changed in 1950 to the *Journal of Projective Techniques*. Gradually, the more objective personality assessment techniques (e.g., the MMPI) were being used and in 1963 the title was changed to the *Journal of Projective Techniques and Personality Assessment*. Projective techniques continued to show disappointing research results and in 1971 this may have prompted the journal's change to its percent title, *Journal of Personality Assessment*. While it is unclear as to what the next change in title will be, it is projected to be something like the "*Journal of Behavior and Personality Assessment*".

Nevertheless, there has remained some doubt as to whether the future direction of assessment will take a distinct behavioral orientation. Even in 1963 when the journal, *Behavior Research and Therapy* made its appearance the issue was raised as to whether there would be a large enough readership to justify its existence (Brady, 1976). However, as Hersen and Bellack (1977) have documented, the future looks very positive as reflected in major journals inaugurated in the United States between the years 1968 and 1970 (Journal of Applied Behavior Analysis, Behavior Therapy, Journal of Behavior Therapy and Experimental Psychiatry). Moreover, several recent behaviorally oriented journals are emerging (e.g., Cognitive Therapy and Research, Biofeedback and Self-control and there are now some specific journals devoted primarily to behavioral assessment (e.g., Behavioral Assessment, Journal of Behavioral Assessment).

Evolution of Nondiscriminatory and Non-biased Assessment Testing as the Context

With the rapid proliferation of tests during the latter part of this century a number of criticisms of tests and testing practices emerged. Much of the controversy has been over tests of so-called "mental ability" or "intelligence" (e.g., Black, 1963; Garcia, 1972; Gross, 1962; Holman & Docter, 1972; Holtzman, 1971; Laosa, 1973b, 1977a, 1977b; Laosa & Oakland, 1974; Martinez, 1972; Mercer, 1972, 1973; Williams, 1971), particularly with minority group children (e.g., Kratochwill et. al., 1980; Reschly, 1979). Indeed, the major controversy in discriminatory or biased testing has been that because minority group individuals typically score lower (or respond differently to questions) on various conventional tests, discriminatory or biased practices will result when vocational and/or educational experiences are denied to these individuals (cf. Laosa, 1977). The argument is advanced that many standardized tests are biased toward people of backgrounds other than that of middle-class, white, and English speaking. While the arguments against traditional testing are not limited to educational settings, it is in educational settings where tests, especially ability measures, have been used to classify individuals for various special education classes.

Biased Assessment Practices in Schools

As noted in the previous section, by 1935 intelligence tests had been applied in a variety of settings. With the development of group tests it became possible to test large numbers of school children (Pintner, 1931). In elementary school:

> The chief practical uses of tests up to the present time have centered around their value for the purpose of classifying children into more or less homogeneous intelligence groups, and also for predicting their future success in school work. These two purposs are intimately bound up with each other. Classification in homogeneous groups is justifiable because intelligence correlates highly with school success, and therefore, the more homogeneous the group the more likely are the children in the group to advance together at about the same rate, be that rate relatively fast, normal, or slow (Pintner, 1931, p. 239).

While homogeneous grouping was widely practiced by 1930 (McClure, 1930), critical reactions to this practice (e.g., Keliher, 1931) as well as negative reviews (e.g. Rankin, 1931) did little to influence the practice that continued well into the future (cf. Carroll, 1978). Indeed, Carroll (1978) notice that research on the efficacy and usefulness of ability grouping had up to 1935 yielded no clear conclusions and a continued negative tone has pervaded the more contemporary period (cf. Svenson, 1962; Findley & Bergan, 1971). Of course, part of the problem has been the inability of research to elucidate solutions to the problem.

Schools have continued to be the focal point for analysis of the use of tests. Unfortunately, testing practices in schools adhered to a very restrictive model, particularly in abiltiy testing:

> Throughout the school system, from elementary school to the university, the use of intelligence tests seems to have been predicated on the assumption that their scores reflected mainly innate or at least relatively unalterable characteristics of students having to do with their capacity to do school work. Although it was noted that average scores were correlated with demographic variables such as socioeconomic class, race, urban/rural enviornment, etc., there does not seem to have been any serious consideration of whether children's home background, or even their schooling, would have any important influence on their performances in mental tests ... the question of whether test scores were biased by cultural factors, for example, was hardly ever raised during the developmental period of the mental testing movement (Carroll, 1978, p. 36).

Over time, issues of bias or discrimination were increasingly raised. Berdie (1965) noted that various tests may lead to discriminatory practices. Mercer (1971, 1973, 1975) supported this observation after studying the relations between membership in ethnic minority groups and placement in classes for the mentally retarded in public schools in California. Mercer (1975) noted:

> We classified every person on the case register into ten groups according to the median value of the housing on the block on which he lived. We found that persons in the lowest socioeconomic categories were greatly over-represented on the register and those

> from higher statuses were underrepresented. When we studied ethnic groups, we found 300 percent more Mexican-Americans and 50 percent more blacks than their counterparts in the general population but only 60 percent as many Anglo-Americans (Caucasians whose primary language is English) as would be expected. Because most Mexican-Americans and blacks in Riverside come from lower socio-economic backgrounds, ethnic group and socioeconomic status are correlated. When we held socioeconomic status constant, Anglos were still underrepresented and Mexican-Americans were still overrepresented in the case register but blacks appeared in their proper proportion (p. 133).

Mercer and her associates also found that this overrepresentation of Mexican-American and black children in classes for the educable mentally retarded was a statewide pattern and not just a local finding. The implication of this was that children from certain low socioeconomic groups or from ethnic minority groups are more vulnerable to being classified as mentally retarded and that certain assessment devices (mainly intelligence tests) are culturally biased.

Of course, these problems are not related to only tests of intelligence and the mental retardation classification. Although IQ tests may not be the primary reason for over and underrepresentation of minorities in special classes (Meyers, Sundstrom, & Yoshida, 1974), legal issues have primarily focused on test bias as the reason for disproportionate representation of minorities in special classes (Reschly, 1979). Thus, it appears that abuses of intelligence testing have received public scrutiny due to social and political consequences, but many of the problems with the intelligence testing have been true of other kinds of norm-referenced assessment (cf. Salvia & Ysseldyke, 1978). Criticisms of standardized assessment have been focused on many dimensions (Laosa, 1973b, 1977; Newland, 1973; Oakland, 1973, 1977; Salvia & Ysseldyke, 1978; Thorndike & Hagan, 1969). Laosa (1977) summarized these criticisms:

1. Standardized tests are biased and unfair to persons from cultural and socioeconomic minorities since most tests reflect largely white, middle-class values and attitudes, and they do not reflect the experiences and the linguistic, cognitive, and other cultural styles and values of minority group persons.
2. Standardized measurement procedures have fostered undemocratic attitudes by their use to form homogeneous classroom groups which severely limit educational, vocational, economic, and other societal opportunities.
3. Sometimes assessments are conducted incompetently by persons who do not understand the culture and language of minority group children and who thus are unable to elicit a level of performance which accurately reflects the child's underlying competence.
4. Testing practices foster expectations that may be damaging by contributing to the self-fulling prophecy with low level achievement for persons who score low on tests.
5. Standardized measurements rigidly shape school curricula and restrict educational change.
6. Norm-referenced measures are not useful for instructional purposes.
7. The limited scope of many standardized tests appraises only a part of the changes in children that schools should be interested in

producing.

8. Standardized testing practices foster a view of human beings as having only innate and fixed abilities and characteristics. (p. 10-11).

These, among other issues, are the primary focus of the remainder of the report.

Summary and Conclusions

In this chapter we have provided an historical perspective on the development of assessment practices in psychology and education. We noted that assessment practices actually have their roots in antiquity. Many assessment practices used today in psychological and educational settings can actually be traced back to activities that occurred hundreds of years ago. Thereafter, developments in France, Germany, England, and the United States formed the basis for developments that would occur in more formal and standardized testing.

A major movement called "differential psychology" formed the basis for the rapid proliferation of ability testing in the United States. Many tests of intelligence were developed to assess children's ability to succeed in school. Many of the tests that were devloped were actually used to place children into special classes or for homogeneous grouping procedures. The early Binet scale and its revisions as well as group tests of intelligence were used for this purpose.

Some of the individuals who were active in development of early tests held views that can be labeled as "racial". Questions were often raised as to the motivations for test development and their subsequent use as a result of those views held. It is clear from an historical perspective that some of these positions could not and would not hold up to empirical analysis. At the heart of many of these early positions was the notion that observed differences in measured intelligence among different racial or ethnic groups was due to genetic differences. This issue has remained a central source of controvery in present day research and writing on test bias.

Major developments also occured in testing personality and behavior. Following WWII, many traditional tests of personality (e.g., Rorschach) were used to assess children and adults. As a movement behavior modification was part reactionary to traditional methods of testing. Developments in this area of psychology and education have had a tremendous impact on both assessment and the nature of special education services provided to school children.

Finally, in the chapter we traced some of the more recent developments in the area of assessment bias in educational settings. Again, it was emphasized that standardized tests of ability have been the primary focus of criticism in research and writing. Unfortunately, many of the issues raised by test supporters and critics alike have not been subjected to empirical research.

Chapter 3

Conceptual Models of Human Functioning: Implications for Assessment Bias

An extraordinary amount of theory and research has been generated that has a bearing on bias in psycholgical and educational assessment. As a result, a tremendous amount of data have accumulated concerning the origins, development, influences, and variations in human behavior. Nevertheless, the wealth of information has clearly not resulted in any integrated view of human performance. Indeed, the current state of knowledge generated from the various conceptual models has not only resulted in the lack of an integrated view of human functioning, but has yielded various conceptual positions that are diametrically opposed.

Because our understanding of human behavior is influenced by basic assumptions concerning the "why" of behavior, assessment practices often become inextricably interwoven with the particular conceptual model of human functioning held by the asseessor. Different models, with their different perspectives of behavior, yield vastly different assessment approaches and data which are used in making decisions relative to classification and intervention. Different conceptual models must be considered in designing nondiscriminatory or non-biased assessment programs (Mercer & Ysseldyke, 1977). Presumably, different models will yield different diagnostic decisions and interventions. The conceptual and psychometric validity and credibility of each particular model must be evaluated and bias examined in light of conceptual and methodological quality within various models.

In this chapter we review seven models of human behavior that influence contemporary assessment practices. The models reviewed include the medical or biogenetic model, intrapsychic disease model, psychoeducational process of test-based model, behavioral model, sociological deviance model, ecological model and pluralistic model. These various models have been discussed by others in the professional literature. For our purposes, these models will also be examined in light of the implications they hold for potential bias in assessment. The models differ in their conceptualization of deviant behavior, assessment procedures and devices (sometimes), as well as the nature of the intervention employed. Because the behavior therapy model has not received as much attention in the nonbiased assessment literature, and because many behavioral procedures such as task analysis, are being advocated in non-biased assessment, we discuss this model in relatively greater detail. Each model is discussed within the context of various components and considerations in its use.

Medical Model

Components

The medical model is one of the oldest approaches guiding assessment and treatment. The medical model can be applied in either a literal or metaphorical context (Phillips, Draguns, & Bartlett, 1975). In this section we view the medical model in its literal sense. That is, abnormal biological systems can be traced to some underlying biological pathology which is then treated. For example, defective hearing (symptom) may be traced to some type of infection (the cause) which may be treated with antibiotics. The prevalance of medical problems in the schools is actually quite high (Schroeder, Teplin, & Schroeder, 1982). For example, May Lau, Lowenstein, Sinnette, Rogers, and Novick (1976) screened 190 second-grade students from two schools in Harlem. They found that 109 (57%) had a total of 170 health problems. A variety of health problems may be found in the school, including those who are chronically ill, those with nutritional disorders (undernutrition, obesity), hearing and visual disorders, dental problems, disorders of bones and joints, infectious disorders, respiratory disorders, allergic disorders, urinary disorders, blood disorders, neurological problems, cardiovascular disorders, as well as drug related problems (Schroeder et al., 1982). It seems clear that a medical model is clearly appropriate to deal with the diversity of medical problems in the schools.

The medical model is a disease-based model. The pathology is assumed to be within the individual. Some theorists consider biological deviations to be the necessary and sufficient factors in the development of the pathology, while others claim that chemical or neurological anomalies are the necessary but not sufficient condition for pathogenesis. Here, environmental conditions may or may not catalyze a constitutional predisposition to pathology.

Considerations

Medical model assessment procedures are clearly justifiable when there is no basis for assuming physiological change in the organism as a result of the socio-cultural environment. Appropriate use of the medical model "should not yield radically or culturally discriminatory results except to the extent that poverty and socioeconomic deprivation are associated with particular groups and elevate the prevalence of poverty-related organic pathologies in these groups" (Mercer & Ysseldyke, 1977, p. 72). Discriminatory practices may very well characterize medical model assessment when they are used to interpret measures of learned behavior (e.g., various forms of disruptive behavior in children, academic skill deficits, etc.). Seventy or more years of biological, biomedical, and genetic research have isolated very few clear physical bases for recognized, psychopathology (cf. Phillips, et al., 1975). While genetic, developmental, neurological and biochemical factors all undoubtedly influence behavior, in reality these factors are not discrete entities. They are interwoven with one another as well as with environmental factors. This may have led Ausubel (1969), in defending the concept of disease to describe abnormal behavior, to contend that it is valid to consider a particular symptom as both a manifestation of disease and a faulty interaction with the environment.

Applications of the medical model may bias assessment in various ways. Organic factors may not always be the cause of an observed medical/physical problem. There is growing recognition that psychological factors may affect a physical condition and that physical symptoms may have no known organic or physiological basis (e.g., DSM-III). In the past, various concepts such as "psychosomatic" or "psychophysiological" have been used to describe the psychological basis for physical or somatic disorders. However, such perspectives may also be of limited usefulness because it implies a simplistic relation between psychological factors and a distinct group of physical disorders when in fact, there may be a complex interaction of biological, environmental, psychological, and social factors contributing to various physical disorders (Siegel, 1983). Lipowski (1977) noted:

> The concept of psychogenis of organic disease...is no longer tenable and has given way to the multiplicity of all disease...the relative contribution of these factors [social and psychological] varies from disease to disease, from person to person, and from one episode of the same disease in the same person to another episode...If the foregoing arguments are accepted then it becomes clear that to distinguish a class of disorders as "psychosomatic disorders" and to propound generalizations about psychosomatic patients is misleading and redundant. Concepts of single causes and cirilinear causal sequences for example from psyche to soma and vise versa are simplistic and obsolete (p. 234).

The point here is that even in the treatment of physical disease, psychological factors may be involved (Melamed & Siegel, 1980). Exclusive reliance on medical assessments may bias treatment in the sense that psychological (or other) aspects of functioning may be involved.

The medical model is being used with increasing frequency in psychology and education. For example, visual and hearing screening are mandated in PL 94-142. A large number of different screening tests are available for assessing physical factors (e.g., Meier, 1975; Conner, Hoover, Horton, Sands, Steinfeld & Wolinsky, 1975; Schroeder et al., 1982). Thus, measures sensitive to organic conditions will be appropriate within the medical model as long as consideration is given to environmental, psychological, and social factors.

As indicated above, problems most often arise when behavioral measures that can be influenced by a variety of environmental circumstances are employed to assess the potential organic origins of a perceived symptom. The more the individual differences observed on a behavioral measure are influenced by environmental factors, the more the measure has the potential of being biased. Such a circumstance may arise when the environmental factors that influence the measure differ across groups. An example of one such measure is the Bender Visual Motor Gestalt Test when it is employed within the medical model to identify potential organic pathology. Although Mercer (1979) employs the Bender in the SOMPA as a measure appropriate for interpretation from within the medical model, she also reports significant correlation between the Bender and various sociocultural measures and between the Bender and ethnic groups. With respect to the latter, when using the Koppitz (1963) scoring system, black children at each age level between 5 and 11, make approximately two errors more than white children. Hispanic children at each of the same age level make approximately one error more than white children. In discussing the influence of social and cultural factors on the Bender, Koppitz (1975) concluded that children from different ethnic groups may develop visual-motor perception skills such as those measured in the test at different rates, and that these differences, in part, may be attributable to factors such as cultural variations in child-rearing practices and the value that varying culture places on these type skills.

Psychodynamic Model

Components

The "psychodynamic model" implies that maladaptive behaviors are symptoms resulting from underlying processes analogous to disease in the literal sense. This model is sometimes labeled the medical model in psychological and psychoeducational practice. Because conceptualization and treatment of abnormal behavior initially resided largely within the domain of medicine, the medical model was extended to treatment of abnormal behavior, both medical and psychological. While the historical developments of the model are not reviewed in detail here, the reader is referred to several sources which discuss this approach (e.g., Alexander & Selesnick, 1968, Kraepelin, 1962).

The psychodynamic approach is characterized by the following: "(a) uses a number of procedures, (b) intended to tap various areas of psychological functioning, (c) both at a conscious and unconscious level, (d) using projective techniques as well as more objective and standardized tests, (e) in both cases, interpretation may rest on symbolic signs as well as scorable responses, (f) with the goal of describing individuals in personalogical rather than normative terms" (Korchin & Schuldberg, 1981, p. 1147). As is evident in the above characterization, the psychodynamic approach is aimed at providing a multifacited description of the client. The psychodynamic approach has also been characterized as involving a great deal of subjective description and inference. This process is said to promote a unique and individual approach to child assessment.

The psychoanalytic model represents one example of the psychodynamic disease paradigm as do many other dynamic models of human functioning. The dynamic approach to assessment of deviant behavior is best elucidated within the context of assumptions held about the internal dynamics of personality (Mischel, 1968). Traditionally, dynamic approaches have inferred some underlying constructs that account for consistency in behavior. Assessment is viewed as a means of identifying some *sign* of these hypothetical constructs which are of central importance in predicting behavior. This *indirect-sign paradigm* in assessment (cf. Mischel, 1972, p. 319) includes a large variety of projective tests (e.g., Rorschach, TAT, Figure Drawings, Sentence Completion Tests) as well as "objective" personality inventories (e.g., MMPI, California Psychological Inventory).

A second feature of the traditional psychodynamic approach is that it assumes that behavior will remain quite stable regardless of the specific environmental or situational context. In this regard test content is of little concern and may even be disguised by making items ambiguous, as is true in projective testing (Goldfried & Sprafkin, 1974). Indeed, a particular response from a projective test is rarely examined in view of the overt qualities of the situation in which it occurred, but is rather interpreted within the context of a complex theoretical structure.

Considerations

Cheney and Morse (1972) have criticized the dynamic approach to assessment on three grounds. One problem is the preoccupation with historical events often in the absence of any verifying data. The second criticism relates to the emphasis during assessment on the individual's presumed unconscious beliefs, attitudes, motivations, and so forth, as interpreted through projections. Cheney and Morse (1972) charge that this technique is bound more in theory than in evidence. Third, behavior is assumed to be a consequence of internalized pathological features. This assumption ignores evidence showing that many behaviors are situational specific.

The use of various psychodynamic indirect measurement procedures has direct implications for non-biased assessment. These measures continue to be used in clinical practice despite data indicating their low predictive validity (cf. Hersen & Barlow, 1976).[1] For example, Golfried and Kent (1972) note that although the interpretation of certain signs on the Bender-Gestalt test (Hutt & Briskin, 1960) has no empirical support (cf. Goldfried & Ingling, 1964; Hutt, 1968), the revised version of the Bender-Gestalt manual presumably discounted these research findings and still recommended the use of questionable interpretations. A rather extensive literature on the comparative (predictive) validity for indirect measurement techniques (Mischel, 1968, 1971) suggests that predictions made on the basis of self-reports are equal to or superior to those made on the basis of indirect measurement techniques that are interpreted and scored by "clincal experts". These findings hold true for a wide variety of content areas (cf. Mischel, 1972).

While there are major problems in the predictive validity of indirect measurement techniques, responses generated in the test situation are also subject to a variety of situational and examiner influences (cf. Hersen & Barlow, 1976). Masling (1960) documented the influence of situational and interpersonal variables and since then a number of writers have further validated this problem (e.g., Hamilton &

Robertson, 1966; Harris & Masling, 1970, Hersen, 1970; Hersen & Greaves, 1971; Marwet & Marcia, 1967; Masling & Harris, 1969; Simkins, 1960; Turner & Coleman, 1962).

Perhaps the most important issue that has been raised over traditional dynamic assessment is its relation to treatment. A number of authers have noted that there appears to be little relation between traditional assessment and treatment (Bandura, 1969; Goldfried & Pomeranz, 1968; Kanfer & Phillips, 1970; Peterson, 1968; Stuart, 1970). Thus, while traditional dynamic assessment may lead to a diagnosis which may in turn lead to the recommendation of a particular treatment, diagnoses resulting from traditional assessment methods cannot accurately predict what particular treatment mode should be implemented (Ciminero, Calhoun, & Adams, 1977; Stuart, 1970).

Psychometric Test-Based or The Psychoeducational Process Model

The psychoeducational process and psychometric test-based model also bear similarity to the psychodynamic disease model in that underlying processes, or specifically process deficits, are said to account for learning and behavior problems. In many respects this model can be considered a part of the dynamic model discussed above. However, in contrast to this model, a psychometric approach is characterized by the use of a variety of individual and group tests to compare individuals along various trait-dimensions. Within trait-theory approaches, various personality structures are said to account for an individual's behavior (Mischel, 1968, 1974). Trait theorists disagree on what traits explain certain patterns of behavior, but generally agree that certain behaviors are consistent across time and settings and that these patterns are expressions or signs of underlying traits.

In contrast to the psychodynamic position, trait assessors typically have placed a high premium on objective administration and scorings of tests. Attempts have usually been made to establish formal reliability and validity of the various measures used. On empirical grounds, this "statistical" approach has proved generally superior to the more "clinical method" in predicting behavior (cf. Korchin & Schuldberg, 1981), but questions have, however, been raised over the manner in which the research reflects the reality of decision making in actual clinical practice.

Closely related to the psychometric approach is the psychoeducational process model. The model can be considered analogous to the psychometric trait model in that assessment focuses on internal deficits, except its context is psychoeducational rather than personality or emotionally oriented. Mercer and Ysseldyke (1977) list six characteristics of this model. These include: (a) the model is a continuous model based upon the degree of deficit present within the child, (b) the model assumes that adequate development of psychoeducational processes are necessary to the adequate development of academic skills, (c) the model is a deficit model, (d) the deficits or disabilities are viewed as existing within the child, (e) deficits can exist unnoticed, and (f) the model is completely culture bound in that processes are considered necessary to the acquisition of socially defined goals (cf. Ysseldyke & Bagnato, 1976).

Within this model exceptionality can be due to one or a combination of three philosophical positions (Quay, 1973). First is the position that exceptional children experience dysfunctions in certain processes that are critical to learning. In this regard, the problem is considered to be within the child and it is assumed "that the dysfunction itself is unremediable and must be bypassed or, at best, be compensated for " (Quay, 1973, p. 166). A second perspective on exceptionality is the experiential defect view in which various dysfunctions (e.g., neurological organization) are due to defects in experience, such as in crawling. A third view is that the child experiences a deficit in which a limited behavioral repertoire is the basis for learning problems. Finally, these appproaches may operate in combination wherever learning problems are due to process dysfunctions, experience defects, and experience deficits (Yesseldyke & Mirken, 1982).

Since a variety of cognitive, perceptual, psycholinguistic, and psychomotor processes or abilities have been cited as causes of children's academic failure, norm-referenced "cognitive" (e.g., WISC-R, McCarthy, Stanford-Binet)," perceptual" (Bender Visual Motor Gestalt Test, Developmental Test of Visual Perception, Developmental Test of Visual-Motor Integration)," psycholinguistic "(e.g., Illinois Test of Pscholinguistic Abilities), and "psychomotor" (e.g., Purdue Perceptual-Motor Survey) tests are used to assess these abilities.

Most of these assessment procedures follow a diagnostic-prescriptive approach. Ysseldyke and Mirkin (1982) note:

> All of the diagnostic-prescriptive approaches based on a process dysfunction viewpoint of the nature of exceptionality operate similarly. When students experience academic difficulties it is presumed that the difficulties are caused by inner process

dysfunctions or disorders. Tests are administered in an effort to identify the specific nature of the within-child disorder that is creating or contributing to learning difficulties. Disorders or deficits are test named (e.g., figure-ground deficiencies, auditory sequential memory deficits, body image problems, eye-hand coordination difficulties, visual association dysfunctions, and mannual expression disorders). Specific interventions are developed to "cure" the underlying causative problems (p. 398).

Considerations

There are several important implications that can be raised with regard to the assessment tactics used within the process or psychometric model. First, since norm-referenced devices are commonly used within the model, the clinician must assume that clients tested have comparable acculturation to those on whom the test was standardized (cf. Newland, 1973; Oakland & Matuszek, 1977). Yet the point has frequently been raised that standardized tests are biased and unfair to individuals from cultural and socioeconomic minorities because they reflect predominantly white, middle-class values and do not reflect experiences and the linguistic, cognitive, and other cultural values and styles of minority individuals (Laosa, 1977). For example, although the norms for some tests (e.g., some group achievement and aptitude tests, the Stanford-Binet, 1972, & WISC-R) are generally good, norming on other instruments are quite inadequate (e.g., ITPA, Leiter International Performance Scale, Slosson Intelligence Test).

A second issue is that research examining components of reliability and validity on various process measures has not been optimistic (cf. Ysseldyke, 1973, 1975, 1977; Ysseldyke & Salvia, 1974; Salvia & Ysseldyke, 1978). For example, several reviews of research on the ITPA (e.g., Bateman, 1965; Buros, 1972; Prager, Cross, & Burger, 1973; Sedlack & Weener, 1973) have drawn attention to these limitations. The magnitude of the problem of inadequate norming, inadequate or incomplete reliability data, or questionable validity is nicely represented in data presented by Salvia and Ysseldyke (1978). Clearly,the potential biased assessment practices is high given the poor psychometric properties of these instruments.

Aside from the psychometric issues of these assessment approaches (i.e., norming, reliability, and validity) an important issue is the degree to which intervention programs based on these assessment models have been effective. A considerable amount of research has been conducted on ability-training approaches (see Ysseldyke & Mirkin, 1982 for a review). These authors noted that there have been major challenges presented to optometric vision training programs (e.g., Keogh, 1974), visual-perceptual training (e.g., Hammill, Goodman, & Wiederholt, 1974), auditory-perceptual training (e.g., Goodman & Hammill, 1973), and psycholinguistic training (e.g., Sedlak & Weener, 1973). Although the jury may still out on these various procedures, there has been considerable compelling evidence that they have not been effective. Therefore, the issue that must be raised is that these procedures may bias the assessment-intervention process. We have labeled this outcome bias (see Chapter 6).

Finally, the approaches based on the measurement of psychological processes rather than by directly observable features raises questions of bias in mental testing (Reynolds, in press). Most of the tests described in this section measure traits or constructs that are not directly observable and are obviously defined differently by different individuals, and are measured on a relative scale. Thus, various criticisms that have been advanced against the trait approach in general(e.g., Kazdin, 1975) would apply to these procedures . For example, one major criticism of trait testing approaches is that the score one obtains on a test is usually thought to reflect the property of the individual assumed to be measured by the test (e.g., intelligence from intelligence tests, visual sequential memory from the ITPA, and aggression from a projective test). As Tyron (1979) has noted, this sets up a test-trait fallacy that begins with the faulty assumption that: (1) test scores are trait measures; (2) trait measures are basic properties of the person; and (3) test scores reflect basic properties of the person. "This sequence essentially converts a dependent variable into an independent variable; hence a measurement is reified into a causal force. It should also be emphasized that the unsound logic of drawing inferences about ability on the basis of observed performance is integral to the test-trait fallacy" (Tyron, 1979, p. 402). It is possible that adoption of this "test-trait fallacy" can lead to bias in assessment.

It may not be useful to lump all tests together and indicate that they are biased simply because they measure processes (Reynolds, in press). Clearly, some tests are better than others on the basis of certain psychometric criteria. However, tests and testors that embrace the process model will continue to have the problem associated with this model as elucidated above (see also Fiske, 1979). Thus, it seems doubtful that addressing the question of bias on a test-by test basis will solve the fundamental problem of the conceptual model embraced by these approaches.

Behavioral Model

Components

Technically, there is no one model of behavior therapy. Also, contemporary behavior therapy, despite commonalities, it characterized by a great deal of diversity. The different approaches in contemporary behavior therapy include applied behavior analysis (e.g., Baer, Wolf, & Risley, 1968; Bijou, 1970), mediational S-R model(e.g., Rachman,1963; Wolpe, 1958) social learning theory (e.g., Bandura, 1969, 1977), and cognitive behavior modification (e.g., Meichenbaum, 1974, 1977; Mahoney, 1974a; Mahoney & Arnkoff, 1978). These approaches are only briefly reviewed here. The reader is referred to Kazdin and Wilson (1978) as well as original sources within each approach for a more detailed presentation. The following section is adapted from Kratochwill (1982).

Applied Behavior Analysis. This form of behavior therapy developed from the experimental analysis of behavior (cf. Day, 1976; Feister & Skinner, 1957; Sidman, 1960; Skinner, 1945, 1953, 1957, 1969, 1974). It emphasizes the analysis of the effects on independent events (variables) on the probability of specific behaviors (responses). Applied behavior analysis focuses on behaviors that are clinically or socially relevant (e.g., various social behaviors, learning disorders, mental retardation, social skills, etc.) and adheres to certain metholological criteria (e.g., experimental analysis, observer agreement on response measures, generalization of therapeutic effects).

Advocates of applied behavior analysis employ a more restrictive use of the term "behavior" than other areas represented in the field of behavior therapy. Behavior refers to "the observable activity of the organism as it moves about, stands still, seizes objects, pushes and pulls, makes sounds, gestures, and so on" (Skinner, 1972a, pp. 260-261). Internal feelings and cognitions are typically not considered a proper focus for the techniques of therapy, research and practice. However, it must be stressed that apllied behavior analysis focuses on the behavior of an individual as a total functioning organism, although there is not always an attempt to observe, measure, and relate all of an organism's response taking place at one time (Bijou, 1976; Bijou & Baer, 1978).

Many intervention procedures associated with applied behavior analysis are derived from basic laboratory operant research [(e.g., positive and negative reinforcement, punishment, time-out, response cost, shaping, fading stimulus control, and many others- see Bijou, 1976; Gelfand & Hartmann, 1975; Kazdin, 1980; Sulzer-Azaroff & Mayer, 1977)]. Assessment emphasizes the individual application of these procedures and a functional evaluation of their effectiveness (Bijou & Grimm, 1975; Emery & Marholin, 1977). Behavior analysis refers to the study of organism-environment interactions in terms of empirical concepts and laws for understanding, predicting, and controlling organism behavior and repeated measurement of a well defined and clearly observable responses (Bijou 1976, Bijou,; Peterson, & Ault, 1968; Bijou, Peterson, Haris, Allen & Johnson, 1969).

Neobehavioristic Mediational S-R Model. The neobehavioristic mediational S-R model is derived from the work of such learning theorists as Pavlov, Guthrie, Hull, Mower, and Muller (e.g., Eysenck, 1960, 1964, Rachman, 1963; Wolpe, 1958). These approaches are characterized by "the application of the principles of conditioning, especially classical conditioning and counter-conditioning to the treatment of abnormal behavior" (Kazdin & Wilson, 1978, p. 3). Although intervening variables and hypothetical constructs play a role in assessment and intervention, covert activities are most commonly defined in terms of a chain of S-R reactions with cognitive formulations de-emphasized.

A number of treatment procedures such as counter-conditioning and systematic desensitization have been used to treat anxiety reactions, phobic patterns, and other strong emotional disorders in children (Morris & Kratochwill, 1983). Systematic desensitization, based on the principle of reciprocal inhibition (Wolpe, 1958), has been successfully used to treat a wide range of child and adult problem behaviors (cf. Bandura 1969; Paul, 1969 b, 1969c; Rachman, 1967; Paul & Bernstein, 1973). Assessment within the mediational S-R model relies on survey schedules (e.g., fear survey schedules) and self-report data, and direct measures of client behavior (as in the use of behavioral avoidance tests).

Cognitive Behavior Therapy. Many of the procedures subsumed under the rubric of cognitive behavior therapy evolved outside the mainstream of behavior therapy (Kendall, 1981). A unifying characteristic of the cognitive behavior therapy approach is an emphasis on cognitive processes and private events as mediators of behavior change. The source of a client's problems are said to be related to their own interpretations and attributions of their behavior, thoughts, images, self-statements, and related processes (Kazdin & Wilson, 1978).

Contemporary cognitive behavior therapy includes Ellis's (1962) rational emotive therapy, Beck's cognitive therapy, and Meichenbaum's self-instructional training. Treatment strategies are quite diverse (cf. Mahoney & Arnkoff, in press, Meichenbaum, 1974, 1977) and include such techniques as problem solving, stress innoculation, self-instructional training, coping skills training, language behavior therapy, thought stopping, and attribution therapy. These techniques represent procedures not generally addressed by other behavior therapy approaches (e.g., applied behavior analysis) and in some cases emphasize components of a given technique where the interpretation for its efficacy is yet to be resolved (Kazdin & Wilson, 1978).

Assessment in cognitive behavior therapy has tended to be quite broad based taking into account many different dimensions of "behavior". Yet, there is still an emphasis on defining the nature of the target problem whether this be overt or covert. In some cases, more traditional functional analysis of behavior which emphasizes a careful examination of environmental antecedents and consequents, as related to a certain response repertoire are explored (e.g., Meichenbaum, 1977).

Some specific purposes for cognitive assessment have been outlined by Kendall (1981):

1. To study the relationships among covert phenomena and their relationship to patterns of behavior and expressions of emotion.

2. To study the role of covert processes in the development of distinct psychopathologies and the behavioral patterns associated with coping.
3. To confirm the effects of treatment
4. To check studies where cognitive factors have either been manipulated or implicated in the effects of the manipulation (pp. 3-4).

Some specific aspects of cognitive behavioral assessment are discussed in Chapter 7.

Social Learning Theory. Social learning theory is based on the work of Bandura and his associates (e.g., Bandura, 1969, 1971, 1974, 1977; Bandura & Walters, 1963) and has evolved considerably over the past few years. Bandura (1974) initially noted that "contrary to the mechanistic metaphors, outcomes (i.e., reinforcing events) change human behavior through the intervening influence of thought" (p. 859). More recently, Bandura (1977b, 1981) has also noted that in addition to outcome expectation, a person's sense of his/her ability to perform a certain behavior mediate performance. Bandura (1977b,1981) refers to these latter expectations as efficacy expectations or self-efficacy, and therefore, suggests they have important implications for intervention. Psychological treatment and methods are hypothesized to produce changes in a person's expectations of self-efficacy, as in the treatment of phobic behavior. Self-efficacy is said to determine the activation and maintenance of behavior strategies for coping with anxiety-eliciting situations. Self-efficacy expectations are also said to be modified by different sources of psychological influence, including performance-based feedback (e.g., participant modeling), vicarious information (e.g., sympolic modeling), and physiological changes (e.g., traditional verbal psychotherapy) (cf. Kazdin & Wilson, 1978).

Intervention procedures such as symbolic modeling (e.g., Bandura, 1971; Rachman, 1976b; Rosenthal & Bandura, 1978), direct modeling (Bandura, 1977; Rosenthal, 1975, 1976) , and self modeling (Brody & Brody, 1977; Micklich & Creer, 1977) have been associated with the social learning theory approach. For example, modeling has been used to treat a variety of children's fears [Moris & Kratochwill, 1983; (e.g., animal fears, inanimate fears, dental and medical fears], socially maladjusted children (e.g., social withdrawal, aggression), distractibility, and severe deficiencies [(e.g., autism, mental retardation) cf. Kirkland & Thelen, 1978)], as well as a wide range of academic behaviors (cf. Zimmerman, 1977).

Social learning theory stresses that human psychological functioning involves a reciprocal interaction between the individual's behavior and the environment in that a client is considered both the agent as well as the target of environmental influence, with assessment focusing on both dimensions of behavior.

Unifying Characteristics. Despite apparent diversity among the different areas within behavior therapy, several dimensions set it apart from traditional forms of psychological assessment and intervention, particularly the test-based psychometric models and psychodynamic models. Contemporary behavior consists of the following characteristics:

(1) a strong commitment to empirical evaluation of treatment and intervention techniques.

(2) a general belief that therapeutic experiences must provide

opportunities to learn adaptive or prosocial behavior.

(3) specification of treatment in operational and, hence, replicable terms.

(4) evaluation of treatment effects through multiple-response modalities, with particular emphasis on overt behavior (Kazdin & Hersen, 1980, p. 287).

Behavior therapy has become very diverse and now includes a number of therapeutic strategies that were once excluded from the field (e.g., rational emotive therapy). Although these characteristics are tied to the therapeutic aspects of the behavioral approach, each can also be conceptually representative of the behavioral approach to assessment. In the sections that follow the methodological and conceptual issues of behavioral assessment are outlined.

Behavior Therapy Approaches to Assessment of Learning and Behavior Disorders.

Behavior assessment has evolved considerably in the past few years. The foundation has been laid for behavioral assessment within the area of social-behavior disorders (e.g., Ciminero et. al., 1977; Ciminero & Drabman, 1978; Hersen & Bellack, 1976) and there has been attention directed toward assessment of learning disorders (e.g., Bijou & Grimm, 1975; Kratochwill, 1982, 1982; Lovitt, 1975a, 1976b, Ross, 1976).

Applications of Behavioral Assessment. Behavioral assessment can be used in treatment, selection, and research (Goldfried & Linehan, 1977). Its application to determine features of the person and environment that maintain deviant behavior is one of the most common uses (see

Bellack & Hersen, 1978; Goldfried & Davison, 1976; Goldfried & Pomeranz, 1968).

Behavioral assessment is also used for selection purposes wherein the results of assessment are used to identify clients for intervention or assign them to diagnostic categories which have relevance to treatment. Behavioral assessment has not been applied extensively in this area (cf. Goldfried & Linehan, 1977; Wiggins, 1973), but there is increasing work in developing measures which allow prediciton of treatment program success.

Finally, behavioral assessment is used in research. Assessment and design methodology have been an identifying feature of behavior therapy and its scientific basis where empirically validated principles and procedures are used for the systematic evaluation of clinical interventions (Bellack & Hersen, 1978). Although this research base has not been limited to one methodology, a large amount research employing behavioral assessment strategies has been conducted through single case experimental methodology, especially in applied behavior analysis (cf. Hersen & Barlow, 1976; Kazdin, 1982; Kratochwill, 1978). A major feature of assessment within behavior therapy single case research is the repeated measurement of the target response (cf. Bijou, et al., 1968; Bijou, et al., 1969), and on cognitive, motor and physiological dimensions (discussed in more detail in Chapter 7). As reflected in the features described by Kazdin and Hersen (1980), an emphasis is placed on direct measurement techniques (actual measures of the target responses through the three content areas), rather than through indirect measurement (e.g., projective tests, perceptual motor scales, personality inventories, etc.).

Some Distinctions Between Behavioral and Traditional Assessment. There are numerous conceptual and methodological differences between behavioral and traditional assessment but the major differences eminate from the underlying assumptions that each approach adheres to in characterizing human performance. It is even possible that the same assessment techniques (e.g., criterion-referenced assessment, direct observation) could be used in both traditional and behavioral assessment. The various ways writers in the behavioral assessment area have contrasted behavioral and traditional [trait (psychometric) or state (dynamic)] approaches to assessment has been compiled by Hartmann, Roper, and Bradford 1979) and is presented in Table 3.1. Behavioral assessment is usually characterized by relatively fewer inferential assumptions about personality, remaining instead closer to observable behavior. As noted in the previous section, most non-behavioral approaches to assessment and treatment conceive of behavior as relatively stable and enduring and relate learning and behavior disorders to internal processes or characteristics.

In behavioral assessment inferred causes of a disorder are bypassed in favor of a careful environmental analysis of the problem and observable skill deficiencies. Moreover, the intervention program would typically focus on specific skill training rather than on underlying process remediation. These approaches have sometimes been identified as a skill training approach (Ysseldyke & Mirkin, 1982). Thus, within a behavioral assessment framework, it is useful to view an

Table 3.1

Table I. Differences Between Behavioral and Traditional Approaches to Assessment

	Behavioral	Traditional
I. Assumptions		
1. Conception of personality	Personality constructs mainly employed to summarize specific behavior patterns, if at all	Personality as a reflection of enduring underlying states or traits
2. Causes of behavior	Maintaining conditions sought in current environment	Intrapsychic or within the individual
II. Implications		
1. Role of behavior	Important as a sample of person's repertoire in specific situation	Behavior assumes importance only insofar as it indexes underlying causes
2. Role of history	Relatively unimportant, except, for example, to provide a retrospective baseline	Crucial in that present conditions seen as a product of the past
3. Consistency of behavior	Behavior thought to be specific to the situation	Behavior expected to be consistent across time and settings
III. Uses of data	To describe target behaviors and maintaining conditions	To describe personality functioning and etiology
	To select the appropriate treatment	To diagnose or classify
	To evaluate and revise treatment	To make prognosis; to predict
IV. Other characteristics		
1. Level of inferences	Low	Medium to high
2. Comparisons	More emphasis on intraindividual or idiographic	More emphasis on interindividual or nomothetic
3. Methods of assessment	More emphasis on direct methods (e.g., observations of behavior in natural environment)	More emphasis on indirect methods (e.g., interviews and self-report)
4. Timing of assessment	More ongoing; prior, during, and after treatment	Pre- and perhaps posttreatment, or strictly to diagnose
5. Scope of assessment	Specific measures and of more variables (e.g., of target behaviors in various situations, of side effects, context, strengths as well as deficiencies)	More global measures (e.g., of cure, or improvement) but only of the individual

(Source: Hartmann, D.P., Roper, B.L., & Bradford, D.C. Some relationships between behavioral and traditional assessment. *Journal of Behavioral Assessment*, 1979, *1*, 3-21. Reproduced by permission).

BEST COPY AVAILABLE

individual's learning as one would view the acquisition of a specific set of skills (and conversely, a learning problem as a set of specific skill deficiencies).

A distinction made by Goodenough (1949) between a "sign" and "sample" approach to test interpretation has often been suggested as another dimension on which to distinguish between traditional and behavioral assessment (Goldfried, 1976; Goldfried & Kent, 1972). When test responses are viewed as a sample, it can be assumed that they parallel the way in which a child is likely to behave in a nontest situation. When test responses are viewed as signs, an inference is made that the performance is an indirect manifestation of some other characteristic. This feature is demonstrated in the previous section wherein we noted that within traditional assessment a child is said to demonstrate low or poor performance on the visual perceptual memory subtest of the ITPA, wherein it is assumed that underlying visual perceptual processes may be impaired. Such an emphasis on sign approaches also promotes determining the underlying causes of academic and/or social problems. Behavioral assessment places less emphasis on historical conditions and so such factors as developmental history is of secondary importance (Haynes, 1978). But when historical factors are considered in behavior analysis, they are examined in terms of interactional history where physical and social conditions result in a wide range of behavioral repertoiries (Bijou, 1976; Bijou & Baer, 1965).

Viewing academic/learning problems within the traditional or behavioral approaches has important implications for test development ("test" is used broadly to refer to a variety of assessment procedures). Goldfried and his associates (Goldfried, 1976; Goldfried & Linehan, 1977; Goldfried & Kent, 1972) provided a conceptual framework for contrasting these opposing models. Within traditional assessment, the nature of the situation in which the individual is functioning is usually of less interest in assessmennt than are such factors as the dynamic or structural components. Within a behavioral orientation the skills conception of learning problems implies that comprehensive and carefully sampled task requirements be reflected within one's assessment. In this context, the conventional notion of *content validity* of the test becomes particularly crucial, since one must obtain a *representative sample* of those situations in which a particular behavior of interest is likely to occur. Thus, in assessment of a learning problem, this includes both the *content* of the test per se, as well as the *situation* in which the test is administered (Bijou, 1976).

Models of Behavioral Assessment. Several models of behavior assessment have evolved and are listed in Table 3.2. These models reflect the diversity that exists in contemporary behavior therapy as well as the movement towards inclusion of more cognitive factors in assessment (Kratochwill, 1982). The basic S-R model was expanded by Lindsley (1964) to include stimulus (S), response (R), contingency (K), and consequence (C). [the "S" refers to the antecedent events or discriminative stimuli, the "R" refers to behaviors, the "K" represents

Table 3.2

Models of Behavioral Assessment

Model	Source
S - R	Ferster (1965)
	Skinner (19)
S - R - K - C	Lindsey (1964)
A - B - C	Stuart (1970)
S - O - R - K - C	Kanfer & Saslow (1969)
	Kanfer & Phillips (1970)
S - O - R - C	Goldfried & Sprafkin (1974)
PI - PA - PI - PE	Bergan (1977)
BASIC - ID	Lazarus (1973)

Source: Kratochwill, T.R. Advances in behavioral assessment. In C.R. Reynolds and T.B. Gutkin (Eds.) Handbook of school psychology. New York: John Wiley & Sons, 1982.

various contingencies (e.g., schedules of reinforecment), and the "C" denotes the consequences of the behavior (e.g., presentation or removal of positive or negative reinforcement)]. An A-B-C (antecedents-behaviors-consequences) model was proposed by Stuart (1970). An even more expanded model was proposed by Kanfer and Saslow (1969) (see also, Kanfer & Phillips, 1970) who added an O to expand to a S-O-R-K-C formulation. Similarly, Goldfried and Sprafkin (1974) presented an expanded S-O-R-C model for a behavioral analysis.

A commonly used model of behavioral assessment is the Kanfer and Saslow (1969) scheme which includes seven specific components:

1. An initial analysis of the problem situation in which the various behaviors that brought the client to treatment are specified;
2. A clarification of the problem situation in which various environmental variables (e.g., stimuli and responses) are specified;
3. A motivational analysis in which reinforcing and punishing stimuli are identified;
4. A developmental analysis in which biological, sociological, and behavioral changes of potential relevance to the treatment are identified;
5. An analysis of self-control in which the situations and behaviors the client can control are identified;
6. An analysis of social situations in which the interpersonal relationships of individuals in the client's environment and their various aversive or reinforcing qualities are specified;
7. An analysis of the social-cultural physical environment in

which normative standards of behavior and the client's opportunities for support are evaluated.

The Kanfer and Saslow (1969) model can assist the professional in clarifying problem behaviors and elucidating environmental factors related to the target problem. Several positive features of this system are apparent (Ciminero & Drabman, 1978). Unlike many systems, the S-O-R-K-C model includes many components ignored by other models (e.g., biological, social-cultural, reinforecment history, developmental factors); the model focuses on positive (assets) as well as negative (deficit) behaviors; in the behavior analysis tradition, the model is individualized for each client, thereby increasing the probability of an individual treatment for each client.

Despite these positive aspects, the Kanfer and Saslow (1969) system should be used in the context of three considerations (Kratochwill, 1982) First, while the model purportedly provides the assessor with a systematic framework for gathering data, the methods for gathering data must be determined somewhat subjectively. Second, the model does not provide a "scientific" approach to interpreting data collected (cf. Dickson, 1975), or selecting an appropriate treatment strategy (cf. Ciminero, 1977). Finally, the model does not provide a model for evaluation of the intervention plan (Ciminero & Drabman, 1978). Thus, the professional must develop a measurement system to evaluate on intervention program.

To address these considerations, systematic research which demonstrates that certain behavioral assessment procedures (e.g., interview, self-report, direct observaton, etc.) are reliable and valid is a high priority (Kratochwill, 1982). In the area of selecting a treatment strategy, efforts are just beginning to select some correct matches (cf. Ciminero, 1977). One promising approach has been presented by Kanfer and Grimm (1977) who proposed a differentiation of controlling variables and behavior deficiencies into categories that can be matched with available intervention strategies. These five categories and some sub-components which are used to organize client complaints presented during an interview assessment are presented in Table 3.3. While the accompanying change procedures are quite general, the "match" can lead the professional into areas where various intervention programs have been quite succesful in the past. With regard to establishing a particular intervention strategy, an evaluation can be conducted through a functional analysis (e.g., Bijou & Peterson, 1971; Bijou & Grimm, 1975; Gardner, 1971; Peterson, 1968). These features would include (1) a systematic observation of the problem behavior to obtain a baseline, (2) systematic observation of the stimulus conditions following and/or preceding the behavior, with a special emphasis on antecedent discriminative cues and consequent reinforcers, (3) experimental manipulation of a treatment which appears functionally related to the problem behavior, and (4) further observation to record behavior changes (Peterson, 1968).

This functional analysis strategy bears similarity to research design procedures, but does not imply that applied research is being conducted. Credible research would require further methodological

Table 3.3

A Behavioral Analysis System

I. Behavioral Deficits

  A. Inadequate Base of Knowledge for Guiding Behavior

  B. Failure to Engage in Acceptable Social Behaviors Due to Skills Deficits

  C. Inability to Supplement or Counter Immediate Environmental Influences and Regulate One's Behavior Through Self-Directing Response

  D. Deficiencies in Self-Reinforcement for Performance

  E. Deficits in Monitoring One's Own Behavior

  F. Inability to Alter Response in Conflict Situations

  G. Limited Behavior Repertoire Due to Restricted Range of Reinforcers

  H. Deficits in Cognitive and/or Motor Behaviors Necessary to Meet the Demands of Daily Living

II. Behavioral Excesses

  A. Conditional Unappropriate Anxiety to Objects or Events

  B. Excessive Self-Observational Activity

III. Problems in Environmental Stimulus Control

  A. Affective Response to Stimulus Objects or Events Leading to Subjective Distress of Unacceptable Behavior

  B. Failure to Offer Support or Opportunities for Behaviors Appropriate in a Different Milieu

  C. Failure to Meet Environmental Demands or Responsibilities arising from Inefficient Organization of time.

IV. Inappropriate Self-Generated Stimulus Control

  A. Self-Descriptions Serving as Cues for Behaviors Leading to Negative Outcomes

Table 3.3

B. Verbal/Symbolic Activity Serving to Cue Inappropriate Behavior

C. Faulty Labeling of Internal Cues

V. Inappropriate Contingency Arrangement

A. Failure of the Environment to Support Appropriate Behavior

B. Environmental Maintenance of Undesirable Behavior

C. Excessive use of Positive Reinforcement for Desirable Behaviors

D. Delivery of Reinforcement Independent of Responding

Source: Kanfer, F.H., & Grimm, L.G. Behavioral analysis: Selecting target behaviors in the interview. Behavior Modification, 1977, 1, 7-28.

requirements such as, but not limited to, a design that eliminates invalidity threats (Kratochwill & Piersel, in press).

Considerations.

There has been a phenomenal amount of writing in the area of behavioral assessment generally but relatively little of this work has focused on theory, research, and practice in bias in child behavioral assessment. Books providing discussion of issues relevant to the assessment of children (e.g., Ciminero, Calhoun & Adams, 1977; Cone & Hawkins, 1977; Herson & Bellack, 1976, 1981) and chapters that focus exclusively on the assessment of children (e.g., Ciminero & Drabman, 1977; Evans & Nelson, 1977; Kratochwill, 1982) have virtually no mention of bias or nondisciminatory assessment.

A number of issues can be raised in child behavioral assessment that have a direrct bearing on assessment bias. A major issue in the field is defining what behavioral assessment is. A reading of the recent literature on behavioral assessment will clearly show that it is remarkably diverse and is becoming even more diverse. A major reason for this is that behavioral assessment is part of the larger domain of behavior therapy which is known to be extraordinarily diverse in theoretical approaches, research methods, and therapy techniques (cf. Kazdin, 1979). Behavioral assessment has always been closely linked with the development of behavior therapy, (Kratochwill, 1982; Mash & Terdal, 1981) (e.g., applied behavior analysis, meadiational S-R approaches, social learning theory, and cognitive behavior modification (Kazdin & Wilson, 1978) . Each of the areas of behavior therapy has tended to include its own assessment techniques and procedures reflective of the theoretical position advanced. Thus, fundamental differences in assessment strategies have occurred across theoretical approaches within the field of behavior therapy. Since behavioral assessment has grown to include such a diverse array of techniques, the study of bias in behavioral assessment becomes a multifacited task which has yet to be properly explicated.

A second and related issue concerns the actual techniques that are to be considered part of behavioral assessment. Bearing in mind these variations in theoretical approaches within behavior therapy, the number of different techniques and procedures subsumed under the rubric of behavioral assessment is growing incredibly large. Defining what behavioral assessment is now and what it will be in the future will likely be determined by these evolving theoretical perspectives rather than a conceptual approach mapping a uniform set of techniques and procedures. Indeed, attempts to define behavioral assessment have typically focused on the theoretical differences between traditional and behavioral approaches (e.g., Hartmann, et. al. 1979; Nelson & Hayes, 1979) with the behavioral perspective being identified by certain conceptual characteristics [(e.g., assessing many modalities, giving primary emphasis to overt behavior, considering assessment a sample of behavior, among others (Kazdin & Hersen, 1980)]. Yet, with different theoretical perspectives on what is to be included within the domain of behavior therapy, attempts to define the field of behavioral assessment will become more difficult when comparisons are made with so-called traditional approaches.

Behavioral assessment has also been said to embrace a conceptual approach that involves a problem solving strategy in the assessment process rather than the use of a set of specific measurement strategies

(Evans & Nelson, 1977; Mash & Terdal, 1981). This conceptual approach may provide some consistency to the field, but expands considerably the range of possible techniques and procedures that can be included in child behavioral assessment. Behavioral assessment has usually been characterized as consisting of some general domains of assessment, including interview, self-report, checklists and rating scales, self-monitoring, analogue assessment, and direct observational measures (cf. Cone, 1978). Yet, it appears clear that expanding frameworks of child behavioral assessment within the context of a problem solving approach allow virtually any technique or test to be considered as child behavioral assessment. For example, traditional projective test formats might be used as they provide information on a client's cognitions or reinforcer preferences or cognitive style. Also, traditional tests could be conceptualized as a format to provide standardized measures of skill performance, such as in the area of IQ testing (e.g., Nelson, 1980) or achievement and perceptual motor tests (e.g., Mash & Terdal, 1981), and neurological assessments (e.g., Goldstein, 1979). Such a diversity of techniques makes it pratically impossible to speak of bias in behavioral assessment in any meaningful way. Rather it would seem more appropriate to address bias in behavioral assessment at a level specific to the type assessment instrument or technique employed.

A third issue that has been the source of activity in the field and has yet to be resolved relates to the psychometric features of behavioral assessment. Many child behavior assessment strategies might be regarded as potentilly biased based on the lack of standardized features in assessment. Possibly due to a rejection of traditional assessment approaches and their associated measurement guidelines, many of the psychometric features of child behavioral assessment (e.g., norming, reliability, validity and generalizability) have not been adequately addressed (Hartmann, et al., 1979; Mash & Terdal, 1981). In the area of norming, for example, concerns have been raised regarding the rather ambiguous meaning of many assessments conducted with children without concrete reference in either a carefully defined population of children or the environment in which they are being assessed. This has raised concerns regarding the appropriate use of data from child behavioral assessments in classification and in establishing specific goals for intervention programs. In addition, there have been relatively few investigations examining the reliability of child behavioral assessment techniques. Behavioral assessors have been primarily concerned with establishing inter-observer agreement on various response measures, but have tended not to establish the reliability of many assessment techniques using conventional psychometric criteria developed for this endeavor. Moreover, the validity of assessment, including such areas as construct, criterion-related and content validation, has many times failed to appear in the field. Although many behavioral assessors have focused on such areas as content validity, the process through which even this has been accomplished has many times been informal, inadequate, and usually incomplete (cf. Hartmann, et al., 1979).

Developing standardized measures has become even more complicated due to controversy over traditional psychometric concepts. Cone (1981), for example, argued that future work in the behavioral assessment field must focus on a paradigm radically different than the traditional psychometric models for establishing reliability, validity, and generalizability. He noted that behavioral assessment procedures are based on a different conceptual model of individual variability than traditional approaches. Thus, the traditional psychometric approaches may be inappropriate for behavioral assessment. As an alternative, he proposed that accuracy be the primary method for establishing the credible psychometric dimensions of future assessment strategies in the field. Based on these issues, it is not at all clear what specific types of psychometric procedures will be established for behavioral assessment techniques or even how the field will deal with devices and procedures that already meet some conventional psychometric criteria.

Regardless of what criteria for establishing the validity of behavioral assessment strategies is finally decided upon, it would seem appropriate that, consistent with the study of traditional test bias, validity criteria be employed as a framework for studying bias. The question posed in the study of bias in behavioral assessment would remain the same as that employed in the study of test bias: Is the assessment procedure equally valid across groups?

A fourth issue that has been a source of some concern in the field relates to the feedback pratitioners have been providing regarding the actual practices of child behavioral assessment in applied settings (e.g., Anderson, Cancelli, & Kratochwill, in press; Swan & MacDonald, 1978; Wade, Baker, & Hartmann, 1979). One finding has been that behavioral assessors have tended to use a great number of traditional assessment devices. Wade et al. (1979) found that nearly half of their respondants who were members of the Association for Advancement of

Behavior Therapy (AABT) used traditional interviews and a large number of projective and objective tests. Such factors as agency requirements for prescribed test use, requirements for testing involving labeling and classification, and a reported difficulty with implementing behavioral assessment in applied settings were offered as possible reasons for this. These results corresponded to other assessment practices of school psychologists reported by Anderson, et al. (in press) who found that select behaviorally-oriented members of the American Psychological Association (APA), Divison 16 and the National Association of School Psychologists (NASP) employed traditional testing procedures and devices. As with the other issues cited above, this issue further complicates the study of bias in behavioral assessment.

The issues raised in the preceding paragraphs convey something of the issues that have been raised in child behavioral assessment. These issues reflect only some of the more general concerns that have emerged but by no means do they represent a comprehensive overview. Nevertheless, the concerns that are now being examined in the field will likely continue to influence both research and practice for some time to come. Hopefully, the issue of bias will receive adequate attention in this area.

Sociological and Ecological Models

Sociological Deviance Model Components. Maladaptive behavior has often been conceptualized within both a medical and social deviance framework. From a medical perspective, maladaptive behavior can be studied in the same way as other forms of illness, while the deviance perspective focuses on maladaptive behavior as the breaking of social rules (Des Jarlais, 1972).

The term "deviance" is a relatively recent one used to describe maladaptive behavior, although the breaking of social rules has been a topic of study for many years (MacMillan, 1977). Terms previously used in reference to this area have included crime, social pathology, and social problems. While there have been several theories of deviance formulated over the years (cf. Des Jarlais, 1972), this section will focus on labeling theory.

During the 1960's a theory of deviance referred to as labeling theory gained popularity. The labeling process is of primary importance within this theory. A major premise within this method is that groups identify with and have different expectations for their conformists and deviants. While conformists are not expected to break social rules (Des Jarlais, 1972), it is also assumed that the expectations and evaluations of others can influence an individual's behavior with regard to following or breaking social rules.

Viewing maladaptive behavior from within a sociological deviance perspective raises at least three questions (Des Jarlais, 1972, 1978; Szasz, 1969).

(1) What behaviors are considered maladaptive and by whom?

(2) What social factors are related to conformity or rule-breaking?

(3) What are the relationships between those who enforce social rules and those who break the rules?

Rule-breaking is generally viewed as deviation from some norm. With reference to social norms, it is not an easy task to articulate exactly what the norm may be for a number of reasons. One problem in defining social norms is that the standards for acceptable behavior often vary across time and geographic location. Whether or not two unmarried adults living together will be viewed as violating some norm, for example, may depend on when the behavior occurs (in 1945 or 1975) and where (a small town or a large urban area).

Another obstacle to clearly defining social norms is that given the same behavior, there may be little agreement on exactly what, if any, norm has been violated. Homosexuality, for example, may be regarded as illegal, immoral, sick, or simply as an alternative sexual preference, depending upon the observer.

Szasz (1969) contended that in attempting to define social norms we can assume only that they consist of psychosocial, legal, and ethical components. Behaviors which are considered to be maladaptive, then, might be thought of as those behaviors which violate some psychosocial, legal, and/or ethical standard.

Labeling theorists have attempted to unravel the relationship between rule-breaking and deviance. Merely breaking rules does not automatically lead to becoming a deviant. Rather, a person must be labeled deviant before the expectancies which activate the deviant role come into play (Des Jarlais, 1972). Examples of this process might include commitment to a penal institution and placement in a self-contained special education classroom.

Lemert (1962) made a distinction between primary and secondary deviance. Primary deviance refers to the initial breaking of social rules, while rule-breaking that occurs after one has been perceived as a rule breaker is termed secondary deviance (not finding employment because of a history of placement in programs for retarded persons is one example of secondary deviance). Other labeling theorists (e.g., Becker, 1963) limited the use of the term deviance to situations in which social expectations for rule-breaking existed.

A deviant label (e.g., mental retardation) does not always follow rule-breaking (e.g., Mercer, 1973). In fact, in the majority of cases the rule-breaker is probably not labeled. Rule-breaking is common to everyone; yet, not all rule-breakers are labeled as deviant. Undoubtedly, there are many individuals who break social rules, but who are not labeled, because their rule-breaking is undetected (e.g., child abusers). In other cases, however, individuals who are known to be rule-breakers may escape the labeling process entirely (Becker, 1963; Scheff, 1966). It is also possible to become labeled without having broken any social rules, through association with or being related to a labeled person, for example (Sever, 1970).

Labeling theory emphasizes the role of those who have the responsibility for enforcing social rules (e.g., the court system, psychologists, teachers, parents). These individuals and groups initiate the labeling process. They have responsibility for deciding who will play deviant roles, that is, who will be punished, treated, or rehabilitated (Des Jarlais, 1972).

Many factors are involved in whether or not those who enforce social rules will or will not confer a deviant label on the rule breaker. Included are the need of the society to have deviant roles filled (e.g., Farber, 1968), the frequency and visability of the rule breaking, the tolerance level of the society for rule breaking (e.g., Szasz, 1969), the social distance between the rule breaker and those who exercise social control, the relative power of the rule-breaker in the system, the amount of conflict between the rule-breakers and agents of social control, and whether or not anyone has a special interest in enforcing penalties against the rule breaker (De Jarlais, 1972, p. 300, 1978). There are also instances in which an individual must be labeled in order to receive services, as occurs under the guidelines of Public Law 94-142.

Deviance in Childhood. Deviance in children has not received as much empirical scrutiny as deviance in adults. Des Jarlais (1978) has pointed out that there are important differences in the study of deviance in children and deviance in adolescents and adults. One major difference is that while adolescents and adults are generally expected to know the social rules and to comply with them, children are not always expected to have developed knowledge of social rules. The study of deviance in childhood, then, focuses upon how children learn the skills and attitudes to follow social rules, or how children become socialized (Gold & Douvan, 1969).

Children in American society are exposed to many different socialization agents. Lippitt (1978) has identified 10 types of socialization agents, all of which attempt to influence the development and values of children:

(1) the schools;

(2) organized religion;

(3) leisure time agencies with recreational, cultural, and character education programs;

(4) the police and courts;

(5) the therapeutic, special correction, and resocialization services (e.g., social workers, counselors, programs for the handicapped);

(6) employment offices and work supervisors of the young;

(7) political leaders who may have an investment in involving the young in political activities;

(8) parents;

(9) peers;

(10) the mass media

At times, there may be different and contradictory definitions or standards of acceptable behavior among and within these groups. Thus, ambivalent expectations for behavior may be imposed on the child, causing stress and often deviant behaviors such as lowered academic performance, hostility, truancy, and withdrawal.

Criticisms of Labeling Theory. Several weaknesses have been detected in the initial work in labeling theory. Some sociologists (e.g., Matza, 1969; Gove, 1970) have noted the relative lack of emphasis given to the role of the rule-breaker in the process of becoming deviant, as compared to the contributions of the agents of social control. Before the dynamics of the labeling process can be fully understood, the actions of both rule-breakers and rule enforcers need to be delineated further.

A second criticism of labeling theory relates to the outcomes of becoming labeled. At issue here is whether or not being labeled necessarily results in negative perceptions and expectancies or in deviant behavior, as is often maintained (e.g., Dunn, 1968). Some authors (see the volume edited by Gove, 1975b) have maintained that the labeling process itself can actually prevent deviant acts by either leading to effective intervention for the labeled person or through a deterrence effect. The response of proponents of the theory (e.g., Becker, 1963; Kitsuse, 1975; Schur, 1975) has been that labeling theory is not an attempt to describe the etiology of deviant acts, but is rather a framework from which to view the actions of all persons involved in situations in which certain behaviors and persons are perceived as norm violators. Conclusions on the outcomes of labeling on the labeled person are mixed because studies in this area are often plagued with methodological problems (Gardner, 1966; Jones, 1973).

A third issue in labeling theory is related to the irreversibility of the labeling process. Once labeled, does a person remain in a deviant role? Robins (1966) presented evidence that most children labeled as deviant become conformists as adults. Gove (1970) also argued against irreversibility, citing the number of mental patients who are released. MacMillan (1977) contended that because different demands and expectations are made in different settings (e.g., school, home), the notion of irreversibility will not always hold. A person viewed as behaviorally disordered in one setting, for example, may not be considered deviant in other settings.

These criticisms have clearly revealed that labeling theory is incomplete, but possibly not invalid (MacMillan, 1977). Future research may very well focus on factors involved in both the direct and indirect effects of becoming labeled, and on racial and cultural biases involved in determining who becomes labeled. The labeling issue is currently receiving attention as part of the debate over present procedures for assessing and classifying children for special education services. The work of Mercer (1970, 1971, 1973) has documented the process whereby some children are identified and classified as mentally retarded, for example, within the public schools. Mercer's work highlighted the disproportionate number of black and spanish-surname children that were being labelled mentally retarded in the Riverside, California public schools thus raising concern over potential bias in the labeling process.

Labeling and categorization, it has often been argued, is a useful way of grouping people who need similar treatment. Yet, categorization or labeling of some people (e.g., handicapped persons) often does not result in academic improvement and can, in fact, lead to negative consequences such as segregation from non-handicappped peers. Clearly, the current labels and categories within the fields of psychology and education need to be further examined (cf. Reynolds & Balow, 1974). In addition, much more work is needed on the direct and indirect outcomes of labeling and mislabeling and on whether or not the outcomes of being labeled in one setting generalize to other settings (MacMillan, 1977). With respect to bias, such research also needs to focus on the possible differential effects labeling may have across groups.

Ecological Model Components. Ecology is the study of the interactions between living organisms and their environment. A more precise definition was provided by Odum (1953), who referred to ecology as the study of the structure and function of nature. Structure includes a description of the living population, including life history, number and distribution of all species in the system, the composition of non-living things, and the conditions under which the population lives.

Function refers to the energy or interaction of organism and environment (Feagans, 1972).

Ecological theorists have attempted to systematically categorize behaviors of species within their environments. In addition, patterns of behavior which account for adaptation or maladaptation to the environment have also been examined (Feagans, 1972).

According to Holman (1977) the field of human ecology is founded in three areas: 1) plant and animal ecology, 2) geography, and 3) studies of the spatial distribution of social phenomena. Holman noted the lack of agreement of basic tenents and principles of the ecological approach among ecologists, primarily because this area is pursued by individuals from many different disciplines and perspectives. Rogers-Warren and Warren (1977) observed that "the meaning of ecology is still evolving" (p. 4) because psychologists, educators, and sociologists who share the term ecology have focused on different aspects of relationships between behavior and environments.

Researchers in the area of human ecology have tried to follow the systematic and precise classification procedures of biology and have evolved methodological procedures to collect and analyze data. This type of research is especially difficult, however, because of the number of variables which must be unraveled in order to examine human environments. Nevertheless, this type of approach can add to our understanding of behavior within a variety of settings.

Ecological psychology is not a new concept. Kurt Lewin used the term "ecological psychology" in a paper published in 1951. He referred to ecology as the interaction between psychological and nonpsychological factors. Later, Roger Baker (1968) used this same definition in an attempt to formulate a theory of behavior settings or ecosystems. Barker and his associates (Barker & Schoggen, 1973; Wicker, 1973; Willems, 1967) tired to describe and classify the various environments or ecosystems of an individual in a systematic way.

Some examples of human ecosystems would include a party, a classroom setting, or a meeting. Barker's work showed that these ecosystems can influence behavior in two ways: Through the physical facilities included in a particular ecosystem, and through the occupants, or human influences present in the behavior setting (Smith, Neisworth, & Greer, 1978).

Proshansky, Ittelson, and Rivling (1970), working in mental hospitals, developed principles about behavior in settings or ecosystems based on Barker's approach. They found that environmental elements such as space, administrative guidelines, furniture, and number of people had a great deal of influence on the behavior of patients. Examples of some of their conclusions are:

> Human behavior in relation to a physical setting is enduring and consistent over time and situation; therefore, characteristic patterns of behavior in a setting can be documented and justified; changes in these characterisitc behavior patterns of a physical setting can be induced by changing the physical, social, or administrative structures which define that setting (Feagans, pp., 1972).

Proshansky et al. (1970) used this approach to modify the behavior of institutionalized patients by changing physical objects in their ecosystem (ward). Gump, Schoggen, and Redl (1963) found that the behavior of a disturbed child differed markedly across physical settings. Rhodes and Tracy (1972) have used an ecosystems approach as one means of conceptualizing the causes of and developing interventions for children with behavioral disorders.

The impact of a behavior setting on behavior has important implications for psychological and educational assessment. The variability of the behaviors of one individual across different settings merits examination. It may be inadequate (i.e., biased) to assess and diagnose behavior in one setting (e.g., school) without observing the same behavior in other settings, with different cues, materials, and people. This point has been repeatedly made by authors addressing the problem of assessing severely handicapped persons (Brown, Nietupski, & Hamre-Nietupski, 1976). It is probably safe to generalize the salience of an ecosystems approach to many different groups.

Assumptions of the Ecological Model. Prieto, Harth, and Swap (in press) pointed out that an ecological perspective of human behavior is based on at least five assumptions about the interaction between an individual and the environment.

(1) Maladaptive or problem behavior does not exist solely within a person but in combination with the ecosystem(s) which the person is an integral part.

According to this assumption, behavior is not the "exclusive property of the child" (Rhodes, 1970, p. 449). Rather, behavior is a result of an interaction or interface between the individual and the environment. Conditions may be present in the environment which can actually elicit disturing behaviors. In addition, indivual(s) in the setting must perceive behaviors as inappropriate. Rhodes (1970) has argued that maladaptive.

This view of disturbance is quite different than the psychodynamic and medical models which locate disturbance within the individual. While a stressful environment may contribute to the problem behavior, no environment or event is significantly stressful in and of itself unless it is interpreted or responded to as such by the person himself/herself.

Several different patterns of faulty interaction between an individual and the environment can be identified. A relatively rare pattern is one in which a person emits inappropriate behaviors in all settings (e.g., self-abusive or self-stimulating behaviors). More commonly, disturbing behaviors occur primarily in only one setting (for example, a child who stutters in a classroom or highly verbal children). A third pattern is that in which problems occur because behaviors which may be adaptive in one setting (an institution) are perceived as maladaptive in another setting (the community).

(2) Ecological interventions must focus on the setting(s) in which the maladaptive behavior occurs.

"The objective is not merely to change or improve the child but to make the total system work" (Hobbs, 1975, p. 114). This assumption requires assessment of the characteristics of the individual, the setting characteristics, and the dissonance between them. Due to the multiple factors involved in such a task, assessment and modification of human behavior in natural settings remains at a primative level of development (Willems, 1969).

(3) Interdisciplinary personnel participate in interventions derived from the ecological model.

Ecological assessment and intervention require an interdisciplinary approach or, "someone who can move freely among and communicate with diverse disciplines in the performance of a liason function" (Hobbs, 1975, p. 120). Teachers, parents, medical personnel, and psychologists often have roles in developing programs within school settings. Effective intervention within community settings, for example, would require the participation of lawyers, economists, employers, and media experts as well.

(4) Ecological interventions must simultaneously focus on many elements of the system.

Willems (1971, 1977) noted the interdependence of ecological networks. Specifically, intervention measures designed to impact change on one element in the system can effect other elements in the system, as well. Modifying a child's behavior in school, for example, can have unintended and sometimes undersirable effects in the home setting. To quote Prieto, et al. (in press), "We can never do merely one thing."

(5) No two individuals and no two settings are the same. This common sense assumption reflects both the strengths and limitations of the ecological model. It is precisely what renders ecological approaches to assessment and intervention so appealing in theory and so difficult to implement in practice.

Ecological Assessment. Ecological approaches to assessment are intended to identify the problems with the interface between a person and his/her setting(s). This is an extremely difficult task, because so many variables are at play. Baer (1977) has stated the problem well: "Assessment of phenomenal reality is an infinite task, like defending against "enemies". We can spend any amount of our resources on it, we will never finish, we will never solve the problem, and if we fail eventually, we will not care much afterward anyway" (p. 116).

Kratochwill and his associates (Petrie, Brown, Piersel, Frinfrock, Schelbe, Le Blanc, & Kratochwill, 1980) presented an ecological framework for school psychologists involved in the implementation of applied behavioral psychology in education settings. Based on Wilem's (1974) discussion of unintended effects in intervention work, these authors presented a conceptual framework for the classification of some types of unintended effects that may occur in behaviors that are not directly manipulated by an individual providing intervention services in educational settings. Some possible types of unintended effects on behaviors are presented in Figure 3.1 In this respect, prediction of 162 (i.e., 3x2x3x3x3) possible kinds of side effects are possible.

As an option in evaluating such side effects (or second-order consequences) Locatis and Gooler (1975) presented guidelines that

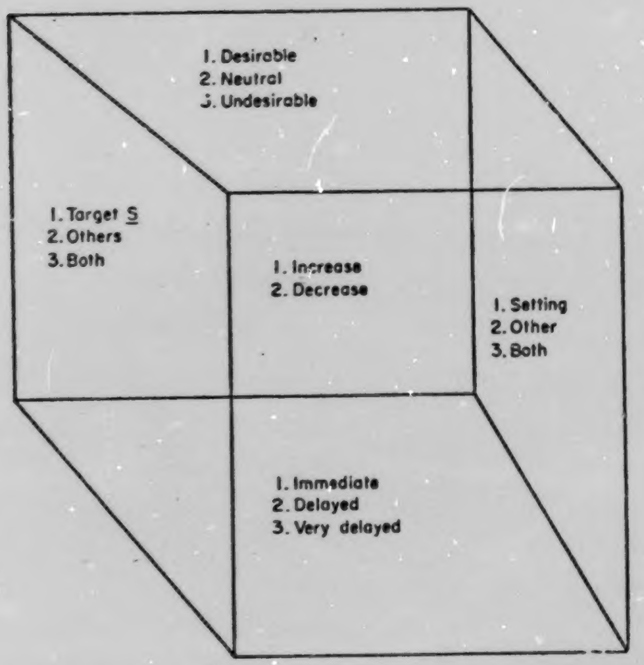
1. Desirable
2. Neutral
3. Undesirable
1. Target S
2. Others
3. Both
1. Increase
2. Decrease
1. Setting
2. Other
3. Both
1. Immediate
2. Delayed
3. Very delayed

Figure 3.1 Classification of some kinds of unintended effects that may occur in behaviors that are not manipulated by the professional. (Source: Petrie, P., Brown, K., Piersel, W.C., Frinfrock, S.R., Schelble, M., LeBlanc, C.P., & Kratochwill, T.R. The school psychologist as behavioral ecologist. Journal of School Psychology, 1980, 18, 222-233 Reproduced by permission).

pertain to individual assessments relevant to educational settings. Fourteen guidelines are presented as a format to evaluate second order consequences (see Table 3.4).

There is no single assessment tool appropriate for evaluating each relevant variable in a given ecosystem. However, there seems to be increasing realization that setting variables must be considered in order to clarify and remediate behavioral and learning problems. A brief discussion of some of these variables and recently developed assessment devices follows.

Assessment of Behavior in Single Settings. Smith, Neisworth, and Greer (1978) and Rogers-Warren (1977) outlined strategies for assessing specific target behaviors and relevant setting characteristics. These include the following:

1. Identify the target behavior: The behavior of concern by name, topography, and function for the subject in the target setting is identified.
2. Assess the physical setting in which the target behavior occurs: Here factors such as instructional space, architectural design, furniture, and physical cues for the target behavior are identified.
3. Assess instructional arrangements: The task here is to evaluate the curriculum content, teaching methods materials, and media.
4. Assess the social situation within the setting: Teacher-teacher, teacher-child, and child-child interactions, reinforcement

Table 3.4 available in Sattler, J.M. _Assessment of children's intelligence and special abilities_ (2nd ed.) Boston: Allyn & Bacon, 1982.

contingencies, and staff time and competence are relevant here.

5. Assess the setting in relation to any existing or anticipated interventions: The question of whether or not the physical and social characteristics which are known to exist in a particular setting will facilitate or hinder a particular intervention in that setting must be considered.

In their text, Smith, et al. (1978) have provided examples of assessment checklists which may be used to assess these types of variables in educational settings.

Moos (1972), working in psychiatric ward settings, developed the Ward Atmosphere Scale (WAS). This device can be used to measure sociocultural aspects of ward environments relative to posthospital outcome. Included on the WAS are measures of patients' involvement in their program, autonomy of patients, order and organization of the ward program, and degree of staff control.

Assessing Behavior Across Settings. Of primary importance here is to determine the effect of different settings on behavior. Assessment devices available for use in single behavior settings are much more common than assessment techniques designed for assessment across settings (Prieto, et al, in press).

Behavior differences across behavioral settings have been observed and described by several researchers (e.g., Gump, Schoggen, & Redl, 1963; Tars & Appleby, 1973). Thomas and Chess (1977) developed interview schedules and behavior checklists which allowed them to compare the perceptions of adults concerning the characteristics of a child's behavior across settings. Their work showed that behaviors which were viewed as problem behaviors in one setting (school) were sometimes not perceived as problems in another (home). They concluded that perceptions of behavior as being appropriate or inappropriate were dependent on the expectations and value systems held by the observer.

Smith, et al., (1978) suggested that an efficient way to change behavior might be to change the setting or to place the person in a different setting rather than attempting to change the person (p. 141). If this is so, attempts to assess behavior accross settings are important and merit more attention and development than have been expended to date.

Assessing Community Factors. It seems logical to assume that there is a strong relationship between the values held by a community and the types of services and programs the community provides. Assessing the effects of community and/or culture on behavior patterns and settings, therefore, becomes relevant within the ecological model. Here, assessment must focus on in-school and out-of-school support services, clusters of settings, and delivery systems through which services are made accessible (Smith, et al, 1978).

In order to assess the role of the community in contributing to maladaptive interactions, it is necessary to study which persons become labeled, how identification occurs, what service delivery systems are available, how they affect the patterns of treatment, and the effectiveness of treatment according to multiple criteria (Prieto, et al, in press). Lewis (1973) pointed out that before we can intervene at the community level, we must establish methods to assess bureaucratic regulations and guidelines which are related to the funding of programs. In addition, methods by which service delivery systems can avoid discord between the individual and his settings need to be examined.

Assessment of community services is, perhaps, the most difficult aspect of ecological assessment and, certainly, the least well developed. There have been, however, some noteworthy efforts in this area. Apter (1977) offered a model for community education based on ecological theory which provides a starting point for assessment of the effectiveness of a community's educational system. Some assumptions of Apter's model are that learning should continue throughout life, facilities should be used efficiently, community participation in educational decision-making should be facilitated, programs which meet the unique needs of children and adults should be provided , personnel should realize that education is not the sole property of any one agency, and research and program development should address the totality of a person's education (p. 368).

Gillespie-Silver (1970) developed a checklist for assessment of local community services. Included were industries, ethnic groups, agencies, funded programs, nonprofit agencies, parent groups, medical, legal, and psychological services and their interactions. A second checklist evaluates services provided by the state and region. Information is also provided about services at the national level. In addition, guidelines arepresented fordeveloping integrated service programs for children. Checklists are provided for use in developing an educational plan for a child utilizing community resources and for developing strategies for implementing the plan which consider the resources and support systems available.

Smith, et al. (1978) presented an example of an inventory which can be useful in assessing community components related to students, services, and professionals. These authors also provide guidelines for comparing student needs with existing as well as unavailable community services for meeting those needs.

Prieto, et al. (in press) concluded that the judicial system is one of the most commonly used means to assess the appropriateness of community services for individuals with problem behaviors. The legislation passed during the last few years (e.g., PL 94-142 and Section 504 of the Rehabilitation Act of 1973) would tend to support this revealing observation.

Consierations in the Use of Sociological and Ecological Models for Non-Biased Assessment The differentiating labels "Sociological" and "ecological" were used in describing the previous two conceptual models of human functioning. Such an apparent distinction may not be viable, however, in practical applications derived from these models because interventionists take both environmental and individual variables into account, although to differing degrees.

The sociological and ecological perspectives on maladaptive behavior both evolved partly in reaction to the restrictions of other models of human behavior. For example, because traditional interventions (e.g., psychotherapy) were typically conducted outside an individual's natural environment, two problems emerged. Any positive changes developed in therapy were not necessarily generalized to other settings. If changes in behavior were generalized, they were not always relevant to other settings. Criticisms of the behavioral model also centered on the generalizability of gains and the narrow focus of intervention. Modifying target behaviors might not be sufficient to alter negative patterns of behavior.

There was also concern about the fairness of viewing the person as the sole problem. Viewing the problem as solely within the person leads to relieving others, including professionals, of taking any significant responsibility for the problem. The sociological and ecological perspectives gained support during the 1960's when the Vietnam War challenged beliefs of what was normal, what was deviant, which behaviors and persons were good and which evil (Prieto, et al, (in press). The arguments generated by the events of those years led many to the conclusion that deviance is relative, that is, deviance depends on the values of the persons making the judgements and the context within which behaviors are viewed. This atmosphere undoubtedly led many professionals to the conclusion that disturbance was created by and assessed in situational contexts, and that effective and ethical treatment required altering those contexts as well as the behavior of the individual.

At the present time there are relatively few formalized systems of assessment based solely on sociological and/or ecological therory. Present assessment approaches influenced by these models of human functioning are eclectic in nature. This is simultaneously a strength and a weakness. The strength lies in the multiple views brought by the expertise of many disciplines. The mere categorization of people of products acquired, which all too often characterizes typical educational and psychological assessment reports, can be avoided. The weakness lies in the lack of any systematic formulation and application of intervention strategies based on sociological/ecological assessment data. It takes, perhaps, less time and energy to focus change efforts on the individual that on the environment in which she/he functions.

More attempts are needed to integrate the insights provided by sociological and ecological models (as well as other models of human functioning) into intervention as well a assessment strategies. In particular, there is a need for (1) instruments and methods for assessment of relevant variables in context and; (2) a technology for assessing the interaction of the selected variables.

The outcome of such efforts could be a more complete and more usable description of behaviors in the context in which they occur that could then be translated into viable intervention strategies. Such an approach to assessment would seem to be consistent with criteria for non-biased assessment.

In addition to the potential contribution of the sociological and ecological models to the development of alternative strategies for colleting non-biased data useful for educational decision making, these models also raise issues that challenge the validity of traditional norm-referenced tests. By highlighting the environmental impact on the way children learn and perform, these models draw our attention to the situation-specificity of many behaviors that we often casually treat as immutable. Given the cultural differences between minority and nonminority children, and consequent potential differences in learning and performance styles, conclusions drawn regarding the non-biased nature of these tests from technical information available to date may be preceived as premature. Continued research from within the scope of these models should yield a better understanding of the generalizability of test data across settings and the potential undesirable by-products of interventions designed from their use, especially as both apply to culturally different children.

Finally, research investigating the assessment issues raised by these methods will help in the design of alternative methods for measuring those constructs presently measured through traditional testing practice. Such instruments would allow for a more effective examination of the construct validity of the test presently employed through the generation of convergent evidence (Campbell & Fiske, 1959).

Pluralistic Model

Components In recent years a pluralistic model has been identified in the literature (Mercer, 1979; Mercer & Ysseldyke, 1977). Technically, this model is more approprpiately a conceptual approach that assists in organizing various assessment strategies that are more responsive to a culturally pluralstic society than any single conceptually derived assessment strategy. Nested within this conceptual approach is an attempt to address the cultural components of the assessment process. Mercer and Ysseldyke (1977) outlines some assumptions of this approach:

> The pluralistic model assumes that the potential for learning is similarly distributed in all racial-ethnic and cultural groups. It assumes that all tests assess what the child has learned about a particular cultural heritage and that all tests are culturally biased. Persons socialized in a cultural heritage similar to those in the test's standardization sample tend to perform better on the test than those not reared in that cultural traditon because of differences in their socialization. A variety of procedures have been designed to estimate the level of performance

> which the child would have achieved if the cultural biases in the testing instrument and procedures were controlled (p. 83).

A number of different measures have been developed that fall within the realm of this conceptual framework. Essentially, those tests and procedures that have been developed as "culture-specific", "culture-fair", or "culture reduced" fall within this domain. Reviews of various instruments that fall into these categories can be found in Jenson (1980) and Sattler (1982) and are reviewed in more detail in Chapter 7 in this volume. For example, the Black Intelligence Test for Children (BITCH) by Williams (1974), the Enchilada Test (Ordiz & Ball, 1972) which has 31 multiple-choice items that deal with experiences common to Mexican-American barrio children, and the test-train-test strategy presented by Budoff (1972) represent some of the more common techniques. Other examples of so-called culture fair tests include the Leiter International Performance Scale, Cattell's Culture-Fair Intelligence Tests, and Raven's Progressive Matrices (see Samuda, 1975 for other examples).

Another set of procedures within this model use multiple normative frameworks for various groups. Although these normative frameworks can be based on local norm-based tests, the most systematic and identifiable strategy in this area is the SOMPA developed by Mercer and Lewis (1978). The SOMPA is actually a system of tests developed to assess children from culturally different backgrounds. The SOMPA does not just represent a pluralistic model, but rather incorporates aspects of a medical model, social system model, and what is being called here a pluralistic model (Mercer, 1979). Sociocultural Scales have been developed within the context of this Pluralistic Model. These scales have the following purpose:

> The Sociocultural Scales determine how much an individual's world differs from the Anglo core culture. Four scales locate an individual ina three dimensional intercept of socioeconomic status, degree of Anglo cultural assimilation, and degree of integration in Anglo social systems. Once an individual's sociocultural group is gauged by the Sociocultural Scales, a normal distribution of WISC-R scores is predicted for that sociocultural group by means of a multiple regression procedure (Figueroa, 1979), p.33).

There has been a considerable amount of material published on the SOMPA and much of this is reviewed in Chapter 7.

Considerations Culture-fair or culture specific tests used within the context of the pluralistic paradigm have been designed to meet the spirit of being non-biased or nondiscriminatory. Generally, such tests have been developed to minimize language, reading skill, speed and other factors that may be culture specific and to minimize cultural differences affecting test content and test taking behaviors (Oakland & Matuszek, 1977). There are, however, several problems with such strategies. To begin with, language is only one dimension on which various tests could be discriminatory. Such factors as social skills, test taking behaviors may even be more important issues. Even if language is a primary concern, nonverbal tests may not be culture-fair because they depend on cognitive behaviors that are related to language systems (Cohen, 1969). Reviews of this literature also suggest that ethnic minorities do not perform any better on so-called culture-fair tests than on more traditional procedures (Arvey, 1972). Nonverbal tests may even be more difficult than verbal tests for certain groups, such as blacks (Sattler, 1982).

Second, there is some concensus that no test can really be regarded as culture-fair (Anastasi, 1961; Vernon, 1965). Moreover, as Sattler (1974) noted, "...no test can be culture-fair if the culture is not fair" (p. 34). Tests can also be ordered on a continuum from highly culture loaded to highly culture reduced (Jensen, 1980). Such tests would differ in the dimensions presented in Table 3.5. As Jensen (1980) notes, changing a test on any one or a combination of these dimensions will not necessarily make the various tests less culturally biased for a certain cultural group. In prediction on a criterion, each test must be empirically examined for bias. However, most tests that can be characterized as culture-reduced have not been subjected to empirical work equivalent to the more common measures used in educational settings (e.g., WISC-R). After reviewing a number of culture-reduced tests, Jensen (1980) concluded:

> None of these attempts to create highly culture-reduced tests, when psychometrically sound, has succeeded in eliminating, or even appreciably reducing, the mean differences between certain subpopulations (races and social classes) in the United States that have been noted to differ markedly on the more conventional

Table 3.5

Dimensions of Cultural Loading on Various Tests

Culture Loaded	Culture Reduced
Paper-and pencil tests	Performance tests
Printed instructions	Oral instructions
Oral instructions	Pantomime instructions
No preliminary practice	Preliminary practice items
Reading required	Purely pictorial
Pictorial(objects)	Abstract figural
Written response	Oral response
Separate answer sheet	Answers written on test itself
Language	Nonlanguage
Speed tests	Power tests
Verbal content	Nonverbal content
Specific factual knowledge	Abstract reasoning
Scholastic Skills	Nonscholastic skills
Recall of past-learned information	Solving novel problems
Content graded from familiar to rare	All item content highly familiar
Difficulty based on rarity of content	Difficulty based on complexity of relation education

Source: Adapted from Jensen, A.R. *Bias in Mental testing*. New York: The Free Press, 1980, p. 637. Reproduced by Permission.

> cultural-loaded tests. On the other hand, some culture-reduced tests show negligible differences between certain widely diverse linguistic, national and cultural groups, which suggests that these tests are indeed capable of measuring general ability across quite wide cultural distances. The fact that such culture-reduced tests do not show smaller mean differences between blacks and whites (in the United States) than do conventional culture loaded IQ tests suggest that the racial difference in test scores is not due to cultural factors per se (p. 713).

Finally, within the SOMPA there is still little evidence that it's use will lead to educational decisions that are not racially or culturally discriminatory (Oakland, 1979). Various criticisms of the SOMPA have been presented in the 1979 School Psychology Digest (Reschly, 1979) and reviewed by Sattler (1982). Because we discuss this assessment procedure in more detail in Chapter 7 , issues are not discussed here. However, it should be emphasized that there is little empirical data to support its use.

Summary and Conclusions

In this chapter we provided an overview of conceptual models of human functioning and their implications for assessment bias. Specifically, in the chapter we reviewed the medical model, intrapsychic disease model, psychoeducational process or test-based model, behavioral model, sociological deviance model, ecological model, and pluralistic model. Each of these models was discussed within the context of its components, assumptions, and features that make it unique and identify it as a separate conceptual framework for work in the assessment field. In addition to this, each model was critiqued within the context of methodological and conceptual issues.

Several major issues need to be taken into account when considering conceptual models of human functioning and their implications for assessment bias. First of all, each model provides somewhat differenct sets of data to be identified in the assessment process. This is important within the context of what aspects of data might be ignored or deemphasized in the assessment process. For example, in many models assessment occurs prior to actual intervention services and therefore does not always address specifically the kinds of outcomes produced once services are identified. Second, a major problem across all conceptual models relates to the lack of research base for many of the theoretical or philosophical features identified. This is a major problem inasmuch as adherence to a certain model might be based more on subjective or philosophical bias than on empirical analysis. Finally, we believe there is some benefit in the future to considering a broader conceptual base for assessment, taking into account each of the different models. Specifically, each of the different models has certain features to assist in the assessment process that another one may not. Thus, individuals assessing children should consider that various models take into account a broader range of conceptual and methodological features to further reduce assessment bias.

Chapter 4

Technical Test Bias

Many individuals have assumed that differences in the mean performance of various groups on tests of cognitive functioning, especially tests of intelligence, automatically connotes bias (e.g., Alley & Foster, 1978; Chinn, 1979; Mercer, 1976; Williams, 1974). This concept of bias assumes that all human populations are essentially equal with respect to their cognitive functioning and that tests, to be nonbiased, should reflect this similarity. As concluded by Alley and Foster (1978), for example, ". . . a test should result in distributions that are statistically equivalent across the groups tested in order for it to be considered nondiscriminatory for those groups" (p. 2).

This concept of test bias has been challenged by many (e.g., Jensen, 1980; Reynolds, 1982). Reynolds (1982) argues that such a position "conveys an inadequate understanding of the psychometric construct [of validity] and issues of bias" (p. 187, parentheses added). Jensen (1980), referring to this concept of test bias as the egalitarian fallacy, calls this position scientifically unwarranted. When such a position is adopted, one removes from the realm of science all chance of empirically determining whether group differences actually exist or are a function of test bias. Group differences could never be studied since any differences found would be by definition, the result of biased measures. Reynolds and Gutkin (1980) point out that ethnic group differences in mental test scores have been a constant and well documented psychological phenomenon. Those holding the above concept of bias would deny the existence of this phenomenon and by necessity conclude that these reported differences are a function of biased measures.

In opposition to those who hold as biased all tests on which performance is associated with group membership, are those who argue that these differences must be examined empirically to determine if findings are a function of bias or real group differences in the measured construct. It should be noted that the empirical study of the validity of tests to determine if group mean differences are real or a function of bias in no way implies racial bias on the part of the researcher. To the contrary, by studying differences among groups, bias is avoided by examining any a priori assumptions regarding possible measured differences. Whether one holds a priori belief that the differences are real or not and the implications one can draw from the findings, are a function of the theory one adopts. This is the nature of science. Researchers in this area are a diverse group of people, some out to validate testing in its present form, others out to reform current practice (Cole, 1981). Regardless of their motives, most are willing to accept science as the arbitor of their disagreements.

Much of the research that has been generated in the study of test bias has employed validation theory to help provide meaning to measured group mean differences. Validity is an estimate of the degree of accuracy with which a test measures what it proports to measure. The more valid a test is, then, the more accurately we can determine if real group differences exist. Given such direction for the study of bias, Alley and Fosters' (1978) conclusion that all groups should have equivalent distributions regardless of the test's validity is a non sequitur. It is valid tests that will tell us if all groups have equivalent distributions.

The Concept of Validity

While there are many who have adopted the study of validity to provide structure to their empirical search for bias in tests, efforts have gone in different directions as a consequence of how the concept of validity has been defined and which aspects of validity are emphasized. Traditional operationalizations of the concept of validity have resulted in validation being segmented into three types: Namely (1) content validity; (2) construct validity; and (3) criterion-related validity.

Content validity is that type of validity that provides information on how well test items sample the content of the domain of behaviors that are expressions of the contruct measured. How accurately scores on a test represent the construct it purports to measure is an issue of construct validity. The third type of validity, criterion-related validity, is established for the purpose of identifying the accuracy in predicting performance in a criterion to which the construct purports to be related.

There are those who suggest that this conceptualization of validity is problematic (e.g., Cronbach, 1980; Messick, 1975). One of the potential dangers of using this tripartite definition is the risk of fragmenting the larger notion of validity to the extent that we lose sight of the more comprehensive picture (Reynolds, in press). The notion of validity cannot be embodied in any single type of validity , regardless of what the use for the test is purported to be. If so embodied, the potential by-product may be the exclusive, or near exclusive, dependence of any one type to establish the validity of a test.

In response to this problem, both Messick (1975) and Cronbach (1980) have encouraged that we bring together all types of validity and recognize them as aspects of one validity, construct validity. These authors suggest that the validity of a test can only be established when one studies various types of information relevant to the accuracy of the test score. Such a conceptualization, encourages diversity in the way we think about, and consequently study, validity. For example, Messick (1980) identifies 17 different types of validity that may be valuable in studying the accuracy of tests. Within such a conceptualization bias in a test is determined by an examination of a variety of evidence all bearing on the construct validity of the test. Evidence of bias in any one area would classify the test as biased, at least on this dimension.

The study of bias through traditional methods, that is, by employing those methods common to the study of content, construct, and criterion-related validity, can be referred to as the study of technical test bias. A large body of literature has emerged in recent years that has studied bias in a technical sense. Researchers in this area differ in how they classify the different efforts. Consistent with the three types of validity traditionally defining the concept, some identify bias in this area as either content bias, construct bias, or criterion-related/predictive bias (Reynolds, 1982). Others choose to classify the three types of validity within two classes of bias (Cole, 1981; Jensen, 1980). The first includes those studies of bias that relate to the use of the test (i.e., related to criteria external to the test). This encompasses those criteria employed in the study of predictive validity. The second entails the study of bias that is internal to the structure of the test. Criteria used in the study of both content and construct validity employed refers to construct bias as it is presently being called.

For the purposes of this report, the latter two-class scheme will be employed. Since the use of the criteria in both classes are employed for the purpose of helping to verify if a test has construct validity we have chosen to call what has more commonly been referred to as "predictive test bias", external construct bias. For the same reason, we refer to the literature that examines the internal structure of a test as internal construct bias.

External Construct Bias

When using predictive validity of criteria in the study of bias, the question of external construct bias relates to how useful the test is in its prediction to some criterion for individuals with differing group membership. Thus, one is not interested in whether or not groups have the same mean score, but if the test predicts the criterion similarly for all individuals, regardless of group membership. Models used to study prediction bias are statistical models most comprehensively based on the linear regression of the criterion variable on the test score. Three major features of the regression system, slopes, intercepts and errors of estimates are often studied.

One of the most comprehensive definitions for this type of bias is offered by Jensen (1980) who writes,

> A test with perfect reliability is a biased

> predictor if there is a statistically significant difference between the major and minor groups in the slope by x, or in the intercepts k, or in the standard error of estimates $SE_{\hat{y}}$ of the regression lines of the two groups. Conversely, an unbiased test with perfect reliability is one for which the major and minor groups do not differ significantly in b, y, x, k, or $SE_{\hat{y}}$ (p. 379).

A circumstance of no external construct bias, therefore, exists when the regression equations for all groups are equivalent. Thus, any prediction to a criterion from a test score would be as accurate for all members of all groups regardless of the score they receive on the measure of the predictor variable. This condition, referred to as homogeneity of regression across groups, simulatenous regression, or fairness in prediction (Reynods, 1982) is depicted in Figure 4.1. Note that in this condition, two individuals from differing groups scoring similarly on the test would receive similar predictions (Y_1, Y_2, or Y_3) regardless of whether or not the pair scored at X_1, X_2, or X_3.

Slope Bias

The slope of a regression line (i.e., the regression coefficient in the regression equation) is the rate of change in the criterion variable as a consequence of a change in the predictor variable. Slope

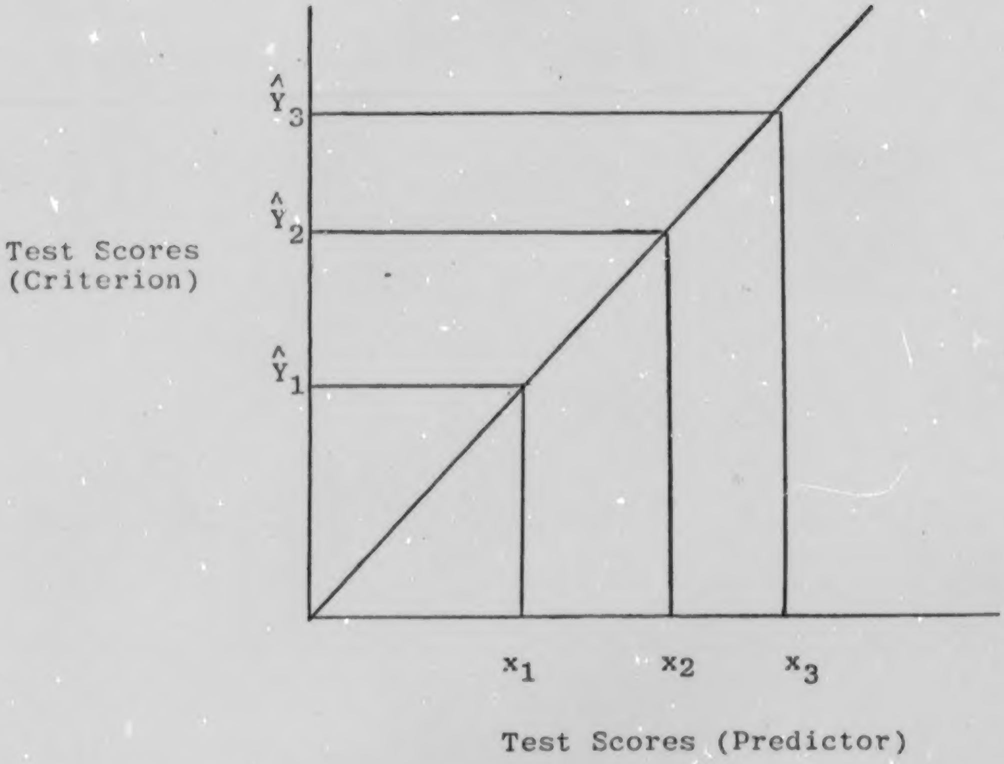

Figure 4.1 - An example of no prediction bias.

bias occurs when the regression coefficients are different for the different groups under investigation; in other words, when the slopes of the regression lines differ. Figure 4.2 graphically depicts an example of slope bias.

As can be seen in the figure, two different regression lines are evident for the different groups. If the regression line for group *A* were used to predict performance on the criterion variable for individuals in both groups *A* and *B*, systematic error (i.e., bias) would occur. For example, if two individuals, one from group *A* and one from group *B*, were to obtain a score of x_1 on the predictor, as can be seen in the figure, if the regression line for group *A* were used, a prediction of a score of Y_1 would not contain systematic error for the member in group *A* and would be an overprediction for the member in group *B*. The more accurate prediction for the group B member (i.e., the one without systematic error) would be Y_1. If the same pair of individuals were to score either x_2 or x_3 on the predictor, the prediction of their scoring Y_1 or Y_2, respectively, on the criterion would be similarly biased for the member of group *B* and unbiased for the group *A* members if the group *A* regression line were used to predict. The more accurate prediction for members in group *B* who score x_2 or x_3 would be Y_2 or Y_3, respectively. Note that in the example depicted, the accuracy of the prediction decreases as a function of increased score.

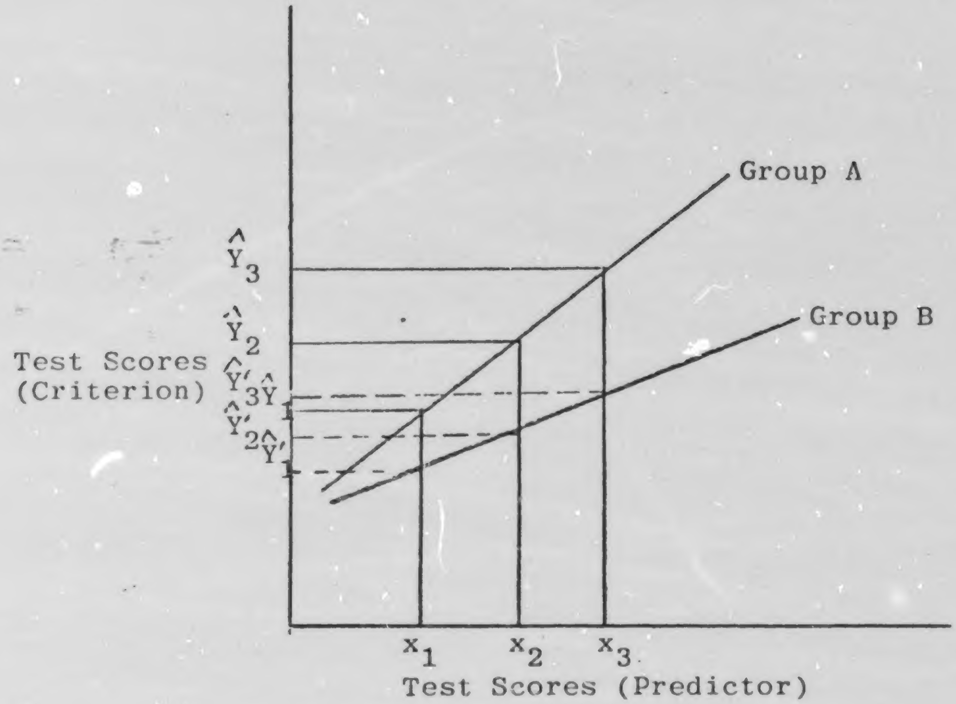

Figure 4.2 - An example of slope bias.

If a single regression line made up of a combination of the regression lines were employed to predict the criterion, the prediction for all members of both groups would be biased. This would be true regardless of the number of individuals in each group. Such is the case if slope bias is evidenced on tests normed on a sample of individuals from groups A and B proportionally selected to represent the makeup, in number, of the total population.

Intercept Bias

Simply stated, the intercept is that point at which the regression line crosses the Y axis. It is the constant in a regression equation and is represented by k differs for different groups. This situation is depicted in Figure 4.3. As can be seen, when such are the circumstances (and there are no differences in slope), the regression lines for the two groups are parallel. If the regression line of one group is used to predict the performance of members from the other group, a constant under- or overprediction will occur. This systematic error, by definition, is test bias. For example, if the group A regression lines were used to predict the performance of our two individuals from the last example (i.e., one from group A and one from group B) and if the pair scored either x_1, x_2 or x_3 on the predictor, a prediction of Y_1, Y_2, or Y_3, respectively, would be made. These predictions would contain no systematic error and be more accurate for the group A member then for the group B member for whom it would be biased. Predictions made for the group B member using the group B

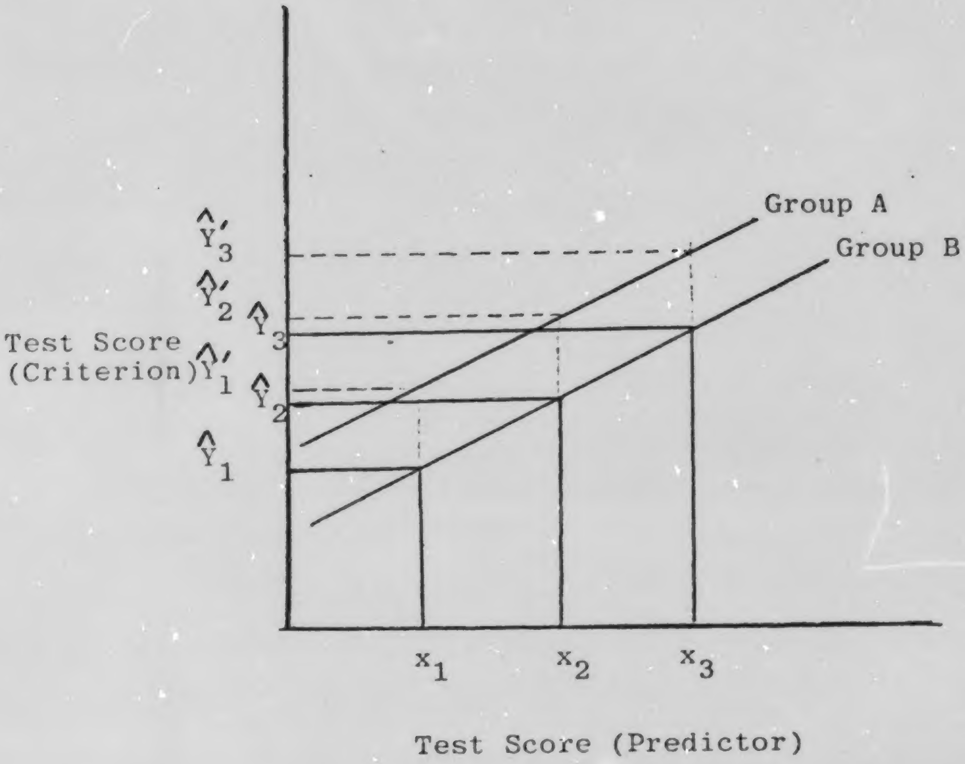

Figure 4.3 - An example of intercept bias

regression line would be Y'_1, Y'_2, or Y'_3, respectively. If a common regression line made up of scores from group A and group B were used to predict, then all predictions made for members of both groups would contain systematic error.

Bias in the Standard Error of Estimate

The third feature of regression that is used as an indicator of external construct bias is the standard error of estimate $SE_{\hat{y}}$. The $SE_{\hat{y}}$ is an index of the amount of error there is in the prediction. Thus, for example, if one plots the scores that are observed on the criterion for a group of individuals all of whom scored x on the predictor and in accordance with the regression line were predicted to have scored Y, a normal distribution of scores around the predicted score Y would be the result. The standard deviation of that distribution is the $SE_{\hat{y}}$. The $SE_{\hat{y}}$ therefore, helps determine the range of potential scores within which one can predict with certain degrees of confidence. If the $SE_{\hat{y}}$ is different for different groups, the test is considered biased. In Figure 4.4 the distribution of estimates for two groups with the same regression lines but different $SE_{\hat{y}}$'s is depicted.

If the observed scores on Y for all the members of group A were plotted, a distribution of errors in estimation would result that would be different than the distribution of errors in estimation plotted for group B. Therefore using the $SE_{\hat{y}}$ for group A to estimate the scores of members of group B would result in a reduced range of estimates and

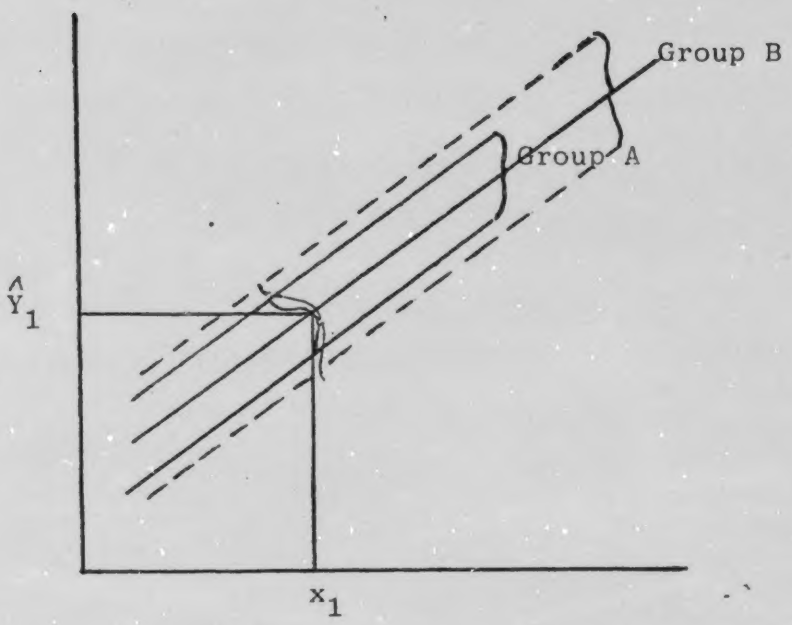
Group B
Group A
$\hat{Y}_1$
x_1

Figure 4.4 - An example of bias in the standard error of estimate.

would be biased. A $SE_{\hat{y}}$ derived from a combination of scores from both groups *A* and *B* would be biased for both groups if, in fact, the $SE_{\hat{y}}$'s are different for each.

Note that in the definition of external bias, systematic error in the predictions for one or more groups on any one of the three factors (i.e. slope, intercept, or $SE_{\hat{y}}$) of regression connotes bias. Of course, bias would also occur if systematic error in prediction resulted from group differences on any combination of the three features. As an example, Figure 4.5 depicts different regression lines for groups *A* and *B* that differ in both slope and intercept. In this circumstance if two individuals, one from group *A* and one from group *B*, were to obtain a score of $\underline{x}_1$ on the predictor, a prediction of $\underline{Y}_1$ on the criterion would contain no systematic error for members of group *A* and would be an underestimation for members if group *B* if the group *A* regression line were used to predict the scores. The prediction for the group *B* member without systematic error would be Y'_1. If two individuals were to score $\underline{x}_2$ on the predictor than regardless of what equation one predicted from, the prediction $_{Y}\underline{2}$ would be as accurate. If both scored $x\underline{3}$ and the regression line for group *A* were used to predict $_{Y}\underline{3}$, the prediction would contain no systematic error for the member of group *A* but would be an overprediction for the group *B* member. The prediction containing no systematic error for the member of group *B* under such conditions would be $Y'\underline{3}$. As can be seen, using either regression equation or a combination of both, would result

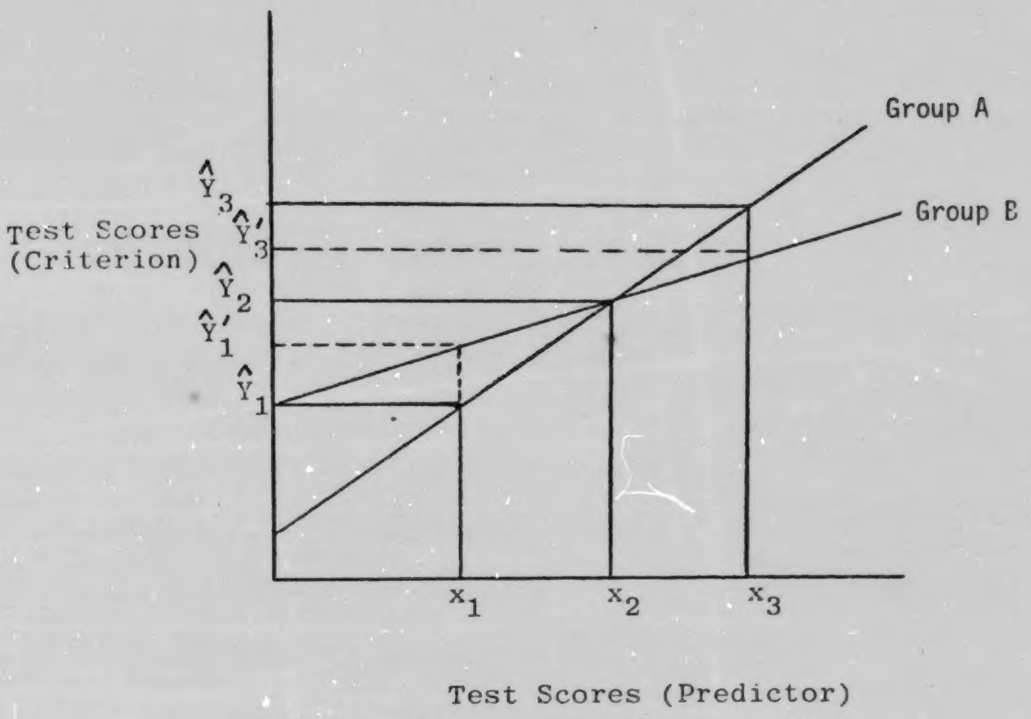

Figure 4.5 - An example of slope and intercept bias.

in systematic error and consequently would be biased.

The various ways one treats a test found to be biased are many. Hunter and Schmidt (1976) argue that automatically eliminating a test that has validity, even though it contains systematic error, may not be the best alternative, especially if you are left with using subjective data to help in decision making. Given the five elemental statistics that can vary between subgroups (i.e., the validity coefficient, the standard deviations of both the predictor and criterion variable, and the reliability coefficients of the predictor and criterion variables), Jensen (1980) argues that each may be examined to determine where bias lies. If not serious, statistical adjustments might be made. Other alternatives would include renorming, or using different tests. These issues, however, are part of the questions regarding fair or unbiased use in the decision-making process and are dealt with in Chapter 6.

Unreliable Tests

One of the unique characteristics of Jensen's definition is its reference to tests with perfect reliability. Linn and Werts (1971) cogently point out that tests without perfect reliability may predict equally well for various groups but would not predict equally well (and thus be biased) if their accuracy was increased by increasing their reliability. Therefore, some conclude that before external construct bias can be established, corrections need to be made in the test scores to account for the unreliability of both the predictor and criterion measures (e.g. Hunter & Schmidt, 1976; Jensen, 1980).

Such a conclusion is important for both theoretical and statistical reasons. From a theoretical point of view, the whole notion of predictive validity is important in that it provides evidence regarding how well the test measures what it purports to measure. As pointed out previously, all types of validity can be viewed as aspects of construct validity. Consequently, a test being judged nonbiased simply because it does not have the reliability necessary to show it to be a biased measure of the construct, would seem to violate the major reason why one would want to measure its relation to a criterion variable in the first place (Jensen, 1980). This argument holds despite the fact that the test's utility in predicting to the criterion variable would remain equally practical across groups.

From a statistical point of view, each of the parameters of interest in the study of external construct bias is sensitive to test reliability. Consequently, just as a biased test may appear to be unbiased due to error in measurement, so too may an unbiased test appear to be biased. "Whatever statistical discriminability a test has, it is only accentuated by improving its reliability" (Jensen, 1980, p. 385). Jensen (1980) describes the potential effects on the interpretation of slope, intercept, and $SE_{\hat{y}}$ bias when either the predictor or criterion variable has less than perfect validity.

When the reliability of the measure of the predictor variable is less than perfect, bias will occur in circumstances where the means of the two groups differ. This will be the case even if the less than perfect reliabilities are equal for the two groups. This latter case will evidence itself in intercept bias. When the reliability of the criterion measure is less than perfect, the $SE_{\hat{y}}$ increases. If the reliabilities across groups differ appreciably on the criterion measure, the outcome will be bias.

As pointed out above, it is recommended that one should first correct for attenuation before concluding that external construct bias exists.

Research on External Construct Bias

Empirical literature in the area of external construct bias has been accumulating rapidly in recent years. While initial efforts focused mainly in the areas of employment selection and college admissions, several studies have recently appeared in the literature relevant to the prediction of school performance.

Studies in the area of external construct bias are potentially fraught with problems. Included among the more serious are: 1) the unreliabilities of the predictor and criterion measures, 2) differing selection criteria for members from the various groups under investigation, 3) inadequate floor or ceilings of tests used for one or all groups studied, 4) inapproriate statistical analysis, and 5) criteria that may reflect differential performance due to experiential factors (e.g., coaching or special training). Complications of these sorts need to be closely kept in mind when evaluating research findings in this area.

Two methods of analysis of group difference have most commonly been used in this literature to lend evidence regarding the potential bias of a test. The first method compares predicitve validity coefficients for different groups while the second examines possible differences in the regression equations derived for the various groups. With respect to the former, only partial information is made available in answering questions of external construct bias when bias is viewed according to the comprehensive definition recommended herein. While it is true that if validity coefficients are different across groups then regression systems must differ, it is also true that if the validity coefficients are the same, it doesn't rule out differences in regressions. In this case, the equivalence of validity coefficients does not rule out external construct bias.

When validity coefficients are used, the way they could be of most value is when a comparison is made _between_ the coefficents to determine the significance of any difference between them. Those investigations that examine separately the validity coefficients for each group to determine if they significantly differ from zero are often erroneous (Humphreys, 1973). In addition to the usual problem of differences in sample size often evidenced in these studies, such a procedure fails to provide empirical evidence regarding the key question as to whether or not the validities among the groups _differ_ from each other. Investigations of this sort have come to be known as single-group validity for one group and not others.

The second method of analysis commonly employed in this area, as mentioned above, involves an analysis of the regression equations of the various groups under investigation. Such an analysis, to encompass the comprehensiveness of Jensen's definition of external construct bias, would have to examine the slope, intercept and $SE_{\hat{y}}$ of the regression systems. Researchers that analyze regressions across groups, as are those that _compare_ validity coefficients, are in search of differential validity. Differential validity refers to a test that has some, yet differing validity for all groups.

Employment Testing. A substantial amount of research has been conducted in the area of employment testing. One of the first major reviews of this literature was published by Boehm (1972). This review examined 13 studies, all involving comparisons between samples of blacks and whites and all involved the study of validity coefficients. Characteristic of most of the studies to follow, the occupations for which the testing was employed ranged in diversity from general maintenance workers to administrative personnel. Of the 160 comparisons of validity coefficients that were made from a total of 57 predictor tests and 38 criterion measures, 4% reported differential validity, a less than chance occurrence ($\underline{p} < .05$). Of the 38 criterion measures, however, most were subjective ratings of job performance. To examine the possibility that there was a difference between the results of those studies employing subjective criterion versus those that employed more objective tests, Schmidt, Bisner, and Hunter (1973) examined 12 of the 13 studies included in the Boehm (1972) review plus seven additional ones. Schmidt et al. (1973) found no difference in the outcomes of the studies when examined according to the subjectivity involved in the criterion measures.

In the Boehm (1972) review, a significant number of studies evidenced single-group validity with tests demonstrating validity for whites and invalidity for blacks. However, since the minority sample sizes were usually smaller than the white sample sizes, these early findings were suspect. Since then, four studies (Boehm ,1977; Katzell & Dyer, 1977; O'Connor, Wexley & Alexander, 1975; Schmidt, Berner & Hunter, 1973), correcting for this error have demonstrated no evidence of single-group validity (Schmidt & Hunter, 1981).

Similarly, some of the earlier studies that found differential validity in higher than chance numbers (e.g., Boehm, 1977; Katzell & Dyer, 1977) have been shown to be methodologically flawed (Hunter &

Schmidt, 1978). Differences in validity coefficients have been demonstrated to be a function of Type I bias resulting from the data preselection techniques used (Schmidt & Hunter, 1981). Avoiding this difficulty, Bartlett, Bobko, Mosier and Hannan, (1978) analyzed 1190 pairs of validity coefficients for groups of blacks and whites and Hunter, Schmidt, and Hunter (1979) examined 712 pairs of validity coefficients for similar group, and both found a less than chance occurrence of significant differences in the comparisons.

In a review of the homogeneity of regression between racial groups in studies done in the employment area, Ruch (1972) examined 20 studies that allowed for the completion of such a reanalysis of the data. The results of the reanalysis had prompted the author to conclude that differential validity occurred at only a chance level of frequency. Citing flaws in the analysis, Jensen (1980) reanalyzed the data reported by Ruch (1972) and concluded that there were no evidence of slope or $SE_{\hat{y}}$ bias. However, his reanalysis identified a highly significant trend of intercept bias. The results consistently suggested higher white than black intercepts with overpredictions for blacks occurring when either a white or common white and black regression equation was used to predict the criterion.

While the studies reported above in the area of employment testing have all focused on black versus white in their comparisons of validity coefficients and regression systems, one study comparing whites and Hispanics has recently been reported in the literature. This study conducted by Schmidt, Pearlman, and Hunter (1980) indicates similar results, that is, no differential validity between groups.

College Adminissions. In the area of college admissions, most studies have looked at potential bias in the Scholastic Aptitude Test (SAT) as it predicts college grade point average. The SAT is a timed multiple-choice paper-and-pencil test that contains two main parts; the verbal part (SAT-V) and the mathematics part (SAT-M). Potential bias in the SAT has been suggested because of the mean difference in the performance of white and various minority groups with the white group scoring higher. The importance of this potential bias comes from the fact that most selective colleges in the United States use the SAT as a criterion for admissions. It should be noted, however, that recent evidence provided by Hardagen (1981) suggests that a wide range of criteria for college admissions is presently used in this country, much more so than in western European countries who depend heavily on admission test results.

As in the employment testing literature, most of studies conducted in this area examine differential validities in the performance of blacks and whites. Those studies examining validity coefficients, on the whole, report no differential validity. For example, in an early study conducted by Stanley and Porter (1967) comparing the validity of the SAT in predicting freshman GPA in three black and 15 predominently white state colleges in Georgia, no differences were found in the validity of the SAT between races when used for this purpose. Yet conclusions from this study must be drawn carefully. A floor effect in the performance of black students on the SAT was found and as Stanley and Porter (1967) indicate, the test was too difficult for approximately one-third of the population of black students. In addition, the study combined heterogeneous samples. It is highly inferential to conclude that the criterion test score (i.e., GPA) means the same thing across institutions. Stanley and Porter (1967) cautiously concluded that the results of their study suggest that the use of SAT scores in predicting freshman GPA was as valid for blacks attending black colleges as whites attending predominently white colleges.

More recent studies comparing regression equations have supported the contention that the use of the SAT in predicting GPA is not biased against blacks when a white or common regression equation is used (e.g., Centra, Linn & Parry, 1970; Cleary, 1968; Davis & Kerner-Hoeg, 1971; Davis & Temp, 1971; Kallengal, 1971; Pfeifer & Sedlacek, 1971; Temp, 1971; Wilson, 1970). To the contrary, a trend in many of these studies indicates that bias, when evidenced, was in favor of blacks (i.e., overpredicted performance on the GPA criterion when using a white or common regression equation).

In a review of the homogeneity of regression of GPA on SAT scores, Linn (1973) concluded that in 22 studies of racially integrated colleges the actual GPA of blacks was overpredicted in 18 of them. In no instance did the SAT underpredict black GPA and in most cases the overprediction was a function of intercept bias. These findings resulted in a panel of the American Psychological Association to conclude that in regular college programs within integrated colleges, with GPA as the criterion, the use of standarized tests for all practical purposes leads to comparable predictions for black and white students (Cleary, Humphreys, Kendrick & Wessman, 1975).

In a review of two studies (Goldman & Richards, 1974; Goldman & Hewitt, 1975) comparing the homogeneity of regressions between white and Mexican-American students using the SAT as the predictor measure and GPA as the criterion measure, Jensen (1980) reports that in both instances the SAT had lower validity for Mexican-American students, and in one study (i.e., Goldman & Richards, 1974) when using a white-derived regression equation, there was a slight tendency to overpredict Mexican-American GPA. This study also indicated that the use of the SAT added little to a prediction made from high school GPA alone.

School Testing. Early efforts to study external construct bias ignored the area of school testing. However, in recent years more attention has been drawn to the use of ability tests to predict academic achievement. Several reasons for this attention have been offered (See Chapter 1), but whatever the reason, several recent research studies have been the result.

Some of the early single-group validity studies reported by Sattler (1974) of individually administered intelligence tests (i.e., the Stanford-Binet and the WISC) supported their validity for samples of black children as well as white children. A more recent single-group validity type study was conducted by Oakland (1979) who reported the validity coefficients of a variety of readiness tests in predicting scores on several achievement measures for groups of black, white, and Mexican-American preschool children from middle and lower SES backgrounds. While no statistic was used to examine differential validity, the size of the coefficients suggested that potential bias may be occurring in the use of readiness tests in predicting non-white performance from white or common regression lines. As pointed out by Reynolds (1982), the lower correlations for non-white groups together with their lower mean criterion score, suggests bias favoring non-whites in predicting early school achievement. However, as discussed previously, single-group validity studies, at most, only allow for inferences across groups.

In a study of the differential validities of seven preschool tests in predicting scores on the Metropolitan Achievement Test (MAT) for samples of black and white preschool children, Reynolds (1978, reported in Reynolds, 1982) conducted an extensive analysis to compare validity coefficents and to examine homogeneity of regression. The MAT was administered one year after the predictor measures. The results of a total of 112 validity coefficients revealed a less than chance number of significant differences. In a study of the 112 regression systems, a significant bias was found across both sex and race with racial bias being significantly more prevalent than sex bias. A further analysis of the data indicated the bias most often occurred in two measures, the Preschool Inventory and the Lee-Clark Reading Readiness Test. The Metropolitan Readiness Test (MRT) showed no bias. When bias occurred it always acted to overpredict black male performance and underpredict white female performance when using a common regression equation.

Several recent studies have appeared in the literature examining the use of the WISC and WISC-R as predictors of scholastic attainment. Much of this research seems to have been spurred by public concern over a disproportionate representation of minority children in classes for the mentally handicapped. Many of these studies have examined the differential validity of these tests in predicting academic achievement as defined in standardized measures such as the MAT. For example, Reschly and Sabers (1979) compared the validity of the WISC-R in predicting performance on the reading and math subtests of the MAT for whites, blacks, Mexican-Americans and Native American Papagos. Consistent with the trend in those studies examined above, an analysis of the regression systems indicated bias resulting in an overprediction of minority performance when a common regression equation was used. The authors found the bias, for the most part, to be a function of differences in intercepts.

Reynolds and Hartlage (1979) predicted reading and arithmetic achievement scores using both the WISC and WISC-R as predictors across samples of blacks and whites. No significant differences were found in predicting achievement for the two groups using different regressions. In a similar study comparing Mexican-American and white children, Reynolds and Gutkin (1980) found the WISC-R performance IQ to differ across groups in its prediction of arithmetic achievement. The difference resulted in an overprediction of arithmetic achievement for Mexican-American children using a common regression equation. There were no differences across groups in the regression equations derived for the WISC-R verbal scale and full scale in the prediction of mathematics, reading and spelling, and the performance scale of the WISC-R in predicting reading and spelling.

In addition to studies examining external construct bias in the WISC and WISC-R, the Stanford-Binet was recently studied (Bonard, Reynolds & Gutkin, 1980) to determine if bias exists in its prediction of academic achievement for black and white children. The results of this study indicated that no systematic bias in prediction was found in an analysis of both the validity coefficients and the regression equations.

While the studies cited above have all been interested in bias in individually administered intelligence tests as it predicts a standardized measure of academic achievement, a few studies have focused on criterion measures that are more global in nature, encompassing criteria reflecting the child's overall performance in his/her role as student. To some, such criteria are viewed as a more valuable standard since it is to this standard that intelligence tests are often asked to predict. Individual tests of intelligence are often used in the schools to help in making decisions regarding special education placement. Clinicians employing the test for this purpose usually infer from its use not only a child's ability to perform on a restricted type of academic task encompassed in a standardized achievement test but also his/her ability to function effectively in the future in his/her role as student. It has likewise been agued that both an IQ test and a standardized measure of academic achievement are measuring the same thing, i.e., the learned ability to master academic type skills and to perform them under standardized conditions (Garcia, 1981). Mercer (19) classified such measures as tests of school functioning and argues that if one wants to use this test as a predictor of school functioning then it should be related to more global criteria than standardized achievement tests. Mercer, therefore, views measures such as the teacher's subjective judgments of the child's ability to perform across the range of subjects as embodied in a report card to be a more useful criterion to predict to when using tests of school functioning (i.e., WISC-R).

While much professional sentiment can be engendered for 1) the hypothesized learned nature of the measured construct and 2) the necessity for examing the utility of the measure for the purpose it is intended (i.e., special education diagnosis and placement), we must conclude that these studies are more appropriately a topic of bias or unfairness in utility and will be considered in Chapter 6. Whether or not one agrees with the nature of the construct or its implied origin, the purpose of the studies reported in this chapter are to provide information on the validity of the tests in measuring constructs. It is hypothesized that these tests measure a construct and that this construct is purported to be related, by definition, to the acquisition of academic skills. To help validate the construct, its relationships to that criterion is an important step.

If the predictor and criterion measures are measuring the same thing and the criterion (i.e., standardized academic tests) measure isn't what its' purported to be, then a better criterion measure should be designed. If the construct is not what it purports to be, then, again, that's a concern that needs to be established empirically through other forms of validity relating to the integrity of the construct. The relationship of the construct to the outcomes of decisions that are made with its use, while considered herein a part of the definition of assessment bias, is not the purpose of the body of literature reviewed in this chapter.

Considerations

As a review of the studies in this section shows, the more recent investigations in all three areas examined, are approaching the study of external construct bias more comprehensively through a comparison of regression equations across groups than earlier studies had done. This more recent trend has encouraged a flurry of investigations over the past decade that have provided rather consistent conclusions. Bias occurs frequently in prediction, and when it does, it usually results in the overprediction of minority performance when a majority or common regression equation is used to predict the criterion. Several explanations have been offered for this bias.

As we discussed previously, one of the effects of unreliable predictor measures is an increase in the difference between intercepts if the mean performance of the groups differ. This would be the case regardless of whether there are differences in the reliabilities or not. We also pointed out in our review that when bias in the regression systems occurred, it was usually of function of intercept bias. For unreliability to account, in part, for the reported bias, the intercept of the minority group would have to be below the intercept of the white group, with overprediction occurring when a white or common regression line is used to predict minority performance. Such is the case in the studies reviewed. Hunter and Schmidt (1976) have suggested that as much as half of such bias witnessed in the literature can be a function of unreliability in the predictor measure. As pointed out above, such error can be statistically eliminated by using estimated true scores in the analysis as opposed to test scores.

If up to half of the bias can be accounted for by test unreliability then what can account for the remaining half? Jensen (1980) points out that this remaining half is the result of "the predictor variable not accounting for enough of the variance in the criterion variable to account for the major-minor groups' mean difference on the criterion" (p. 514). Jensen (1980) also concludes that there are, at present, no explanations for the phenomenon but he postulates, along with others, that it may be a function of differences in the criterion measure caused by such factors as achievement motivation, interests, work and study habits, and personality traits affecting persistence, emotional stability, and self-confidence. Any one or combination of these may influence the criterion measure.

We would like to further point out that these same factors may also account for differences in mean performance on the predictor across groups as well. If such were the case and these factors were uncorrelated to the predictor and influenced the criterion to a similar degree, then no external construct bias would be evident. If they influenced the criterion measure more than the predictor, then bias in the direction observed in the above studies would result if the influence of such factors were negative.

It should also be noted that if factors such as those identified above did influence the predictor measure, it would have serious implications for the construct validity of the test. Additionally, if these factors were irrelavent to the predictor, it would not necessarily be identified in the study of the internal structure of the test during an examination of internal construct bias. The reasons for this will become apparent in the next section when we take a look at methods employed to study internal construct bias. In addition, further light will be shed on this problem when we discuss alternative approaches to the study of construct validity in Chapter 5.

Internal Construct Bias

When looking at the external evidence of construct bias as we did in the last section, the test for which we were interested in judging bias was examined as it predicted to some criteria to which it was purported to be related. When examining evidence of internal construct bias, no such external criteria are used. Rather, the response of groups are examined to determine if differences across groups are evidenced in the structure of the response patterns or in the specific items that make up the test. Criteria of the sort traditionally employed in evaluating content and construct validity are used in making judgments regarding internal construct bias.

Sometimes the criteria traditionally used to judge construct validity are external (e.g., another construct measure). Thus, there is often quite a bit of similarity between criteria of this sort and that criteria employed in examining external construct bias. Conceptually, this is accounted for by the fact that predictive validity is but one aspect of construct validity. Whether or not investigations employing external criteria are included under the traditional heading of predictive or construct validity is, to a certain extent, arbitrary. An often made distinction is that the criteria used to establish predictive validity are a set of behaviors more often acquired by one who possesses more of the construct than less. In addition, the behaviors are usually identified as more specific in nature and perceived to serve more of a practical purpose in determining the usefulness of the test. The use of academic skills as criteria to validate a measure of intelligence is one example.

The external criteria used in those investigations purported to lend evidence to construct validity are usually more general and less practical in nature. For example, the validity of one construct is often partly accomplished by relating it to another construct to which it is hypothesized to be related.

For the purposes of the present review, this distinction is not made. The study of bias as evidenced by differences in group performance on an external criteria are reviewed under the heading of external construct bias. Traditional construct validation studies employing factor analytic procedures are included as part of internal construct bias.

The reason for this distinction is in keeping with our perception of this purpose of traditional validation procedures. From this perspective, the value of traditional validation procedures is in telling us how well a test measures a construct, not its usefulness in decision making. The question as to whether or not a test is useful in decision making is very complex, much more so than can be answered with predictive validity studies of a traditional variety. Therefore, the study of technical test bias is presently perceived as answering questions related to how well the test measures the construct across groups, not how effective it is in predicting outcomes of complex decisions. Since the value of any construct rests ultimately in its use and cannot be divorced from this purpose, we also include in our perception of bias those practices that result in differential treatment across groups. This discussion will be taken up under the topic of outcome bias in Chapter 6.

In the remainder of the present chapter, a number of common statistics generated from test responses are examined to determine the presence of internal construct bias. Evidence of bias in any one of the areas reviewed provides grounds for labeling a test suspect. Examined in the following sections are internal consistency bias, factor structure bias, and item bias. The last of these biases, item bias, is examined according to the four methods most popularly employed in its analyses; the group x items interactions method, the item response theory method, the distractor analysis method, and the judgmental method. In addition to these issues of internal construct bias, a brief discussion of what is called "facial bias" in the literature will be included.

Internal Consistency Bias

One statistic of a test that can be examined across groups to determine if there is evidence of internal construct bias is the intercorrelations among the test's items. This statistic is a reflection of the internal consistency of a test and in measurement terms is one indication of the test's reliability. If the groups under investigation each evidence a high reliability coefficient then what is being measured is being done so with high accuracy for the groups and no bias is suggested. A discrepancy in the reliability coefficient between two groups would suggest either (1) the items are more difficult for the group with the lower estimate or ,(2) the item intercorrelations are different or, (3) both item difficulty and item intercorrelations explain the difference. If differences in the internal consistency estimates exist, it would therefore be necessary to find out if they can be accounted for by item difficulty since such differences could be the result of items being truly more difficult for one group than another. If such were the case, then the test would not be biased. If item difficulty was not the reason for differences in the estimates, then such differences can be attributed to the item intercorrelations. In this case, the test would be considered biased as it would suggest the possibility that the test is not measuring the same thing across groups. An internal consistency reliability coefficient of approximately .90 is an arbitrary standard often used to connote a good test in this regard.

With respect to the WISC-R, two recent large scale investigations of internal connstency were conducted by Oakland and Feigenbaum (1979) and Sandoval (1979). In the Sandoval study, internal consistency estimates for the various WISC-R subtests[3] were computed for over 1000 whites, blacks and Mexican-Americans. Variations in estimates across groups ranged upward to only .04 except for the object assembly subtest. On this subtest differences in the internal consistency estimates were .16 between whites and blacks and .20 between Mexican-Americans and blacks with blacks having higher estimates in both comparisons.

Similar findings are reported in the Oakland and Feigenbaum (1979) study using similar groups. Differences in internal consistency reliability estimates ranged upward to .06 for all but the object assembly subtest. In the Oakland and Feigenbaum study, the internal consistency reliability on the object assembly test for whites was higher than for Mexican-Americans and blacks. The estimates for whites, blacks and Mexican-Americans were .74, .64, and .67, respectively.

With respect to other tests, Jensen (1974) reports estimates of internal consistency for the Peabody Picture Vocabulary Test (PPVT) and the Raven's Colored Progressive Matrices for similar groups. Estimates for his samples on these tests were also similar. On the PPVT estimates ranged from .95 to .97 across groups on the Raven, between .86 and .91. For whites, blacks, and Mexican-Americans, Green (1972) reports similarly high and consistent results ranging from between .90 and .92 for scores on the California Achievement Test.

In addition to the internal consistency estimates on the WISC-R, Oakland and Feigenbaum (1979) also report estimates for the Bender-Gestalt Test for blacks, whites and Mexican-Americans. These estimates suggested similar internal consistencies ranging from a low of .72 for Mexican-Americans to a high of .84 for whites.

In another study, Dean (1977) reported on the internal consistency of the WISC-R for Mexican-American children who had been tested by white examiners. In comparing these estimates to those reported in the predominently white standardization sample of the WISC-R by Wechsler (1974), Dean found, albeit higher, similar and consistent estimates.

From the evidence reported to date, there appears to be no marked group differences in the average degree of accuracy in measuring whatever the test measures. As pointed out earlier, however, we can only infer from this statistic that the same construct is being measured across groups. To be precise, this measure tells us that the tests are measuring the same thing accurately; whether they are measuring the same or different constructs accurately is not determined from this statistic. In addition, this indicator does not tell us if an irrelevant factor or factors are differentially affecting performance across groups even if it is measuring the same construct.

Factor Structure Bias

Evidence bearing on the first of these issues (i.e., whether or not the same construct is being measured across groups) is provided by the most common of construct validation techniques, factor analysis. Sometimes referred to as factorial validity in validation studies, this technique identifies clusters of items or subtests that correlate highly with each other. It also identifies those that don't fit in the clusters. When employed in the study of internal construct bias, the patterns of these interrelationships are studied across groups for congruence. It is reasoned that tests that have different factor structure may be measuring different psychological occurrences in response to the test items. Tests that have similar factor structure are not considered biased with respect to there criteria.

A number of techniques have been proposed to compare factor analytic findings across groups. Some are capable of testing for statistically significant differences between groups (Jensen, 1980; Joreskog, 1971), while others look for similarities in the results (Harman, 1976; Katzenmeyer & Stenner, 1977).

In a reanalysis of Nichols' (1972) data that originally reported the intercorrelations among 13 tests (seven of which were subtests of the WISC) for large samples of white and black 7 year-old children, Jensen (1980) found no significant differences across groups for the g factor loadings (first principle components of the factor analysis) extracted from the intercorrelations.[4]

In a reanalysis of data presented by Mercer and Smith (1972), Jensen (1980) again found no significant differences among white, black, and Mexican-American children of ages 7 and 10 years old on the g factor from eleven WISC subtests. However, the two factor solution using a varimax rotation of the principle components in each of the three groups yielded unclear results. The verbal and performance factors that emerged provided mixed results that Jensen attributed to sampling error (as few as 48 subjects were used for one of the groups).

However, other factor analytic studies of the WISC provide much clearer evidence of similarity of factor structure for the WISC across groups when the two factor solutions were compared (e.g., Silverstein, 1973).

Similar evidence has been found when the test scores on the WISC-R of various groups were factor analyzed. In comparing the factor structure of the WISC-R across whites, blacks, Mexican-Americans and Native American Papagos, Reschly (1978) found substantial congruence when the two factor solutions were compared. When using the three factor solutions (the three factors comprise Wechsler's Verbal and Performance scales minus the Coding, Arithmetic and Picture Completion subtest which make up the third factor that is often referred to as the freedom from distractability factor), Reschly also found congruence between the white and Mexican-American samples.

A further analysis of Reschly's (1978) data by Jensen (1980) showed no significant differences among the four groups on the principle component. Similar evidence continues to appear in the literature regarding the similarity of the factor structure of the WISC-R regardless of the factor analytic procedure used, the statistic employed to study factor structures' similarities or differences, the characteristics of the sample (i.e., normal or referred) and the memberships of the groups that are studied (Blaka, Wallbrown & Engin, 1975; Dean, 1979; DeFries, Vandenberg, McClearn, Kuse, Welson, Ashton & Johnson, 1974; Gutkin & Reynolds, 1980; Oakland & Feigenbaum, 1979; Vance, Huelsman & Wherry, 1976; Vance & Wallbrown, 1978; Wallbrown, Blaka, Wallbrown, Engin, 1975; Wallbrown, Blaka & Wherry, 1973, 1974). Not only has this finding been consistent with regard to the WISC-R but similarly with (1) the WPPSI across blacks and whites (Kaufmann &

Hollenbeck, 1974; Wallbrown, Blaka & Wherry, 1973); (2) the MRT across blacks and whites (Reynolds, 1979); (3) the McCarthy Scales of Children's Abilities across blacks and whites (Kaufmann & DiCuio, 1975); and (4) the Goodenough-Harris Drawing Test across blacks, whites, Mexican-Americans, and Native American Indians (Merz, 1970). The sum of these findings leads to similar inferences regarding the constructs the various tests purport to measure; that is, the constructs are the same across groups.

While these findings provide strong confirmatory evidence that the constructs these tests are measuring are the same across groups, they do not indicate that the constructs are being measured to the same degree. Irrelevant factors that are not intended to be included in the measure, yet are present for either one or all groups, are not detectable through the use of factor analysis procedures.

Item Bias

While those studies that have been reported so far under the heading of internal construct bias have all examined the overall pattern of responses to identify the integrity of a measures construct validity across groups, none have dealt with the analysis of specific items or series of items in a search for evidence of bias. The search for item bias, the oldest practice among all in trying to ferret out bias in testing (see, for example, Eells, Davis, Havighurst, Herrick, & Tyler, 1951), is primarily designed to ensure that the individual items used in a test, contribute equally to the meaning of what is measured across groups. The two most popular methods for identifying biased items are the group x item interaction method and the item-response theory method.

Group x Item Interaction Method. Approaches within the group x item interaction method usually apply either analysis of variance of correlational techniques to the study of item bias. When studying item bias through the use of an analysis of variance design, the group x item interaction term is of major interest (Cardall & Coffman, 1964; Cleary & Hilton, 1968). Such an interaction is an indication that the items are exacting in different ways for different groups. If such were the case, it could be concluded that the items may not mean the same thing for the various groups under examination. A similar effect can be noticed by correlating item difficulties for different groups. The correlation between the rank order of item difficulties or decrements across groups will be low if the items are biased (Jensen, 1976). Once such biasing effects are found, it then becomes necessary to pinpoint those items that are the most biased ones. A variety of procedures have been offered to conduct such analyses (Angoff & Ford, 1973; Angoff & Sharon, 1974; Veale & Foreman, 1975).

The use of the ANOVA technique for identifying item bias has met with the identification of only small percentages of performance variance being accounted for by the group x item interaction. For example, approximately 2% to 5% of the variance in WISC-R performance is accounted for by this interaction when the responses to items of black and white children are compared (Jense, 1976; Miele, 1979).

Correlational procedures likewise have produced evidence of little item bias. For example, rank order of item difficulties across groups has resulted in consistently high correlation. Rank order correlations of item difficulty among white, black and Mexican-American samples of children for the Raven's Progressive Matrices, PPVT, and WISC-R have all been of the .95 to .99 magnitude (Jensen, 1974, 1976; Sandoval, 1979). A similarly high rank order correlation is reported for item diffculties between black and white children on the Stanford-Binet Intelligence Scale (Jensen, 1976).

Jensen (1976) also proposed a rank order correlation of the decrements instead of the item difficulties to identify bias in items. A decrement is the difference in the difficulty indices of two adjacent items when the items have been ranked for difficulty within groups. Such procedures have produced correlations in the upper .90's for the Raven's across white, blacks, and Mexican-American comparisons, respectively, for the WISC-R (Sandoval, 1979). Lower rank order correlations of decrements ranging from .65 to .79 were found across groups (black, white and Mexican-American) for the PPVT.

If there has been any common finding in the item x group interaction studies other than the fact that the bias found appears to be very small (2% - 5% of the variance), it is that the more unreliable the item (usually the more ambiguous items) the more chance the item will turn up biased. One thing that has not been found in these studies is a consistent theme in the content of biased items. As Flaugher (1978) pointed out, it was the early hope of researchers in this area that such themes could be identified and systematically eliminated from tests. However, such has not been the case when we judge item bias using group x item inter-interaction approaches. The only thing that is accomplished by expounding a test of its biased items is to make the test more difficult for all groups since such items tend to be the moderate to easy items in the test (Flaugher & Schroder, 1978).

Item-Response-Theory-Method. Approaches for detecting item bias within the item-response theory method are relatively new and statistically sophisticated. These approaches all use item-response theory for identifying differences in item characteristic curves across groups. An item characteristic curve depicts the relationship of the ability level of the test taken with the probability of a correct response. If the same construct is being measured for all groups studied then one would expect this relationship to show no differences. Various item-response theory models can be applied for this purpose ranging from a more complex three parameter logistic model (Lord, 1977, 1980) to a simpler application of the Rasch model (Durovic, Note 1; Wright, Mead & Draba, Note 2). Lord (1977) argues that the item-response theory method is more appropriate than the item x group interaction method for detecting item bias.

While research using item-response theory approaches have typically yielded results similar to the group x item research, biased items with interperable themes have recently been identified (Cole, 1981). Scheuneman (1979), for example, reports that negatively worded and unfamiliar format items appear biased against black youngsters. Cole (1981) concludes that further research in this area is needed given the only moderate agreement on which items in any given test are biased across samples and which are biased in the same samples when different methods are used.

Distractor-Analysis-Method. A third method for identifying bias in items is through the analysis of distractors. Distractors are the incorrect responses provided as possible alternatives in items employing a multiple-choice format. For a test to be unbiased with respect to its distractors, the incorrect alternative should have the same relative degree of attractiveness across groups. In a study of two popular multiple-choice intelligence tests, the PPVT and the Raven's Progressive Matrices, Jensen (1974) found errors on the PPVT to be nonrandomly distributed on many items; but unexpectedly he also found significant differences between blacks and whites in their choices of responses on 26% of the items. On the Raven's, a significant difference in the type error in choice of distractors was found between blacks and whites on 12% of the items.

Further analysis, however, revealed this difference to be determined not as a function of item difficulty but rather to be age-related. In an analysis of the groups' responses to the Raven's, all cases where potential bias was evidenced, showed that the black children responded similar to white children approximately two years younger. When white children in the third and fourth grades were compared to fifth and sixth grade black children, the difference in choice of distractor largely disappeared. Jensen (1974) concluded, therefore, that the systematic error is consistent with our understanding of the underlying construct and consequently not evidence of bias.

Judgmental Method. One final method for identifying bias in items will be briefly mentioned. This method typically employs the use of expert judges in detecting item bias. While this method has been adopted in recent years by several test developers, its practice has never been empirically justified. To the contrary, the procedure has been demonstrated in several investigations to be no better than detecting test bias than through random selection (Jensen, 1976; Plake, 1979;

Sandoval & Miille, 1979). In the Sandoval and Miille (1979) study, for example, groups of black, white, and Mexican-American college students were asked to judge WISC-R items to determine how easy the items would be for white, black and Mexican-American children. The results indicated that none of the groups of judges were able to identify which items were those that had been empirically determined to be either more difficult for blacks and Mexican-Americans or of equal difficulty for all children. Similar results were found by Jensen (1976) when expert judges were employed.

Facial Bias

The use of judges to determine a type of bias referred to as facial bias has been proposed in recent years (e.g., Anastasi, 1976; Cole & Nitko, 1981). This type bias should not be confused with those efforts discussed above to help identify item bias. Judgmental methods for detecting item bias, when employed, are done so to improve the validity of a test and thus help reduce bias. The notion of facial bias has nothing to do with the validity of a test. Rather it is a form of bias in that it offends certain groups of people or creates a perception of validity-based bias. Cole and Nitko (1981) note:

> Facial bias would occur when particular words or item formats appear to disfavor some group whether or not they, in fact, have that effect. Thus, an instrument using the male pronoun "he" throughout or involving only male figures in the items would be facially biased whether or not such uses affected the scores of women.

The examination of a test by judges to remove items containing facial bias are usually employed for socio-political purposes or for purposes of principle or values (Cole, 1981).

Summary and Conclusion

Evidence from a review of the studies of internal construct bias lead to the overall conclusion that little if any bias can be found in the internal structure of many of the popular psychological tests commonly employed in decision making. However, this conclusion may be premature with respect to item bias. As noted, given the inconsistency of the results across and within the various approaches for detecting item bias, more research is needed to determine if themes can be identified that typically bias sets of items for one or more groups. Although recent findings point to this possibility, the prospects for findings of any practical significance does not look very promising. Given our earlier review of external construct bias combined with our present review of internal construct bias, one would have to conclude that the search for what Cole (1981) calls a "bias bombshell" has just not turned up anything; and prospects for its location in the future while restricting our hunt to the parameters defined in the technical test bias literature will likely provide us with more of the same information.

One question regarding the validity of tests in measuring their respective constructs that is not answered by the technical test bias literature is whether or not differential group mean performance is influenced by factors that are irrelevant to the construct being measured. As pointed out previously, if motivational or emotional factors, for example, influence performance differently across groups then marked differences can result in group means and go undetected in this literature. The criteria examined above has informed us that the use of these tests in predicting certain limited criteria do not appear to be biased against minority group members and may in fact be biased in their favor. In addition, the literature has also told us that in most cases the tests appear to be measuring the same constructs with a high degree of accuracy. What the literature does not tell us is if there are any situational factors such as self-confidence, achievement motivation and the like, that are differently influencing the measure of that same construct across groups. This topic is taken up in the next chapter.

Chapter 5

Situational Bias in Psychological Assessment

Elaboration of the concept of assessment bias must take into account factors that extend beyond the "internal" issues discussed in previous chapters. In this chapter, we review some factors that have been conceptualized as potentially contributing to assessment bias resulting from factors in the external testing situation. These factors are conceptualized as separate from the items per se, and include the following areas: (1) test-wiseness (e.g., practice efforts, coaching), (2) sex of examiner, (3) race of examiner, (4) language factors, (5) expectancy effects, 6) motivational factors, (7) situational considerations, and (8) scoring considerations. Each of these areas are discussed within the context of methodological and conceptual issues raised in the area. Also, some areas are specified for future empirical work in the assessment area. Finally, some tentative recommendations are advanced for practice in the area, especially with regard to psychological and educational assessment of children.

Test-Sophistication

Test-sophistication or test-awareness refers to a potential source of bias when different persons participating in testing have different amounts of coaching or practice prior to taking the test. A number of authors have discussed issues in this area (e.g., Anastasi, 1981; Jensen, 1980; Messick, 1981). Test sophistication is not at all a straight forward concept. Indeed, issues in this area are characterized by considerable controversy. Research in the general domain has focused on both practice and coaching (or training) effects. Jensen (1980) has discussed both of these areas and has advanced some summary statements that are reviewed here.

Practice

Jensen (1980) defined practice as "taking the same or similar tests two or more times at various intervals, without any implication of special instructions or specific coaching in test taking" (p. 590). Based upon other literature in this area (e.g., Jarvis, 1953; Vernon, 1938, 1954a, 1954b, 1960; Wiseman & Wrigley, 1953; Yates, 1953), he advanced 12 conclusions on the effects of practice (Jensen, 1980, pp. 590-591):

1. Practice effects are natrually greatest for naive subjects, that is, those who have not been tested before.
2. Retesting of naive subjects on the identical test, after a short interval, shows gains of about 2 to 8 IQ points for various tests, averaging about 5 IQ points. (Regardless of the tests used in the various studies reviewed here, gains are converted to a scale with = 15, which is the usual for IQ.)
3. There is considerable variability in practice effects among individuals. Bright subjects tend to gain more from practice than dull subjects.
4. The curve of practice gains is very negatively

accelerated with repeated practice; that is, there are rapidly diminishing returns of repeated practice on the same or similar tests, yet slight gains have been shown on up to five repetitions, beyond which there is no further improvement. Practice gain between the first and second test experience is usally as great or greater than the total of all further gains from subsequent practice trials.

5. For naive subjects, age makes little difference in the amount of practice effects. There are more examples of large practice effects in young children, however, simply because fewer of them than of older children or adults have had prior experience with tests.
6. Practice effects differ, on the average, for various types of tests, showing the smallest gains for information, vocabulary, and verbal tests generally and the largest gains for nonverbal and performance tests, probably because the materials of the latter tests are less familiar to most subjects than are verbal and informational questions.
7. Practice effects are greater for tests comprised of heterogeneous types of items than for homogeneous tests.
8. Practice effects are about 10 to 25 percent less for untimed tests than for speeded tests.
9. For naive subjects, practice gains are greater on group-administered paper-and-pencil tests than on individually administered tests.

10. Practice effects show surprisingly little "transfer of training," with the gradient of practice gains falling off steeply from identical tests to parallel forms, to similar tests, to different types of tests. The average practice gain from Form M to Form L of the Stanford-Binet, for example, is only 2 or 3 points. Parallel forms of groups tests show average practice gains of 3 to 4 points after one practice session and 5 to 6 points after several practice sessions. One large-scale study showed a total gain of 6 IQ points over the course of eight parallel forms given to London school children (Watts, Pidgeon, & Yates, 1952).
11. Practice effects are not appreciably diminished by improving the usual test instructions or by giving short practice tests on easy items prior to the actual test. There seems to be no substitute for taking an actual test under normal test conditions for a practice effect to be manifested.
12. The practice effect is quite lasting; about three quarters of the gain found after one week is maintainend up to six months, and half remains after one year.

The conclusion reached on the basis of work in this area has been that practice contributes very little to bias in individual or group differences in test performance. However, whether or not this can be an acceptable general conclusion is debatable. Several issues in this regard are advanced at the end of this section.

Coaching

In contrast to practice, coaching refers to an active attempt to teach test taking skills, broadly conceived. Within this framework would be included such activities as providing instructions in how to take the test, modeling actual or similar items, providing strategies to formulate a response, and providing feedback on the tester's performance. Actually, any type of intervention could be implemented within the coaching paradigm.

Again, Jensen (1980) has advanced several conclusions from empirical work in this area:

1. Coaching is quite ineffective unless accompanied by practice at taking complete tests under regular test conditions. According to Vernon, the leading expert on the topic, "coaching without practice is singularly ineffective, regardless of how protracted it is" (1960, p. 131).
2. The typical gain from several hours of coaching plus practice gain on a similar test is about 9 IQ points, or a coaching gain of 4 or 5 points over and above the gain due solely to the practice effect of taking a similar test once or twice previously.
3. The coaching effect is greatest for naive subjects and diminishes with prior test-taking experience. Even with equal prior testing experience, there are substantial individual differences in gains from coaching; the

moderately bright tend to gain most.

4. Coaching gains are greater on nonverbal and performance-type items than on verbal and information items. Also, numerical reasoning and arithmetic problems are more susceptible to coaching gains than are items based on verbal knowledge and reasoning.
5. Age and sex show no consistent interaction with coaching effects.
6. The effects of coaching are highly specific, with little transfer to other types of tests, and at times there is even negative transfer to dissimilar tests.
7. The maximum effects of coaching are achieved quickly; further gain does not result from coaching prolonged beyond the first few hours. One study found three hours to be optimal.
8. A study of educationally disadvantaged children in Israel found that coaching on a nonverbal intelligence test substantially improved the test's validity, that is, correlation with teachers' marks and with the Verbal IQ of the WISC (Ortar, 1960).
9. The effects of coaching seem to fade considerably faster than the effects of practice per se. A study by Greene (1928) shows the decline over time in the gains on Stanford-Binet IQ from coaching children on the very same test items or on similar items; the control children were tested at the same times as the experimenntal groups, but they were never coached, and

so their gains represent only practice effects. (Jensen, pp. 591-592).

Test Sophistication and Interaction with Race and Social Class

The effects of practice and coaching have been investigated on both racial and social class dimensions (e.g., Baughman & Dahlsdrom, 1968; Dugin, Osborn, & Winick, 1969; Dyer, 1970; Costello, 1970; Turner, Hall, & Grimmett, 1973). These studies, reviewed by Jensen (1980), generally show negative results on the interaction of practice and coaching effects with either race or social class. However, some small effects were noted in the Dyer (1970) study, but this was conducted on a college sample. Thus, from this limited literature one can conclude that minimal effects of practice and coaching may appear.

Considerations

Test sophistication represents an important variable in any testing and ultimately influences the validity of the tests employed. Before any firm conclusions can be made in this area it is important to consider several issues, including the conceptualization of test sophistication and the potential areas that can be used for intervention in this area. To begin with, test sophistication does not refer to a set of homogeneous factors. The nature of what test sophistication is or what effects it has cannot be addressed in the abstract, but rather depends on the specific features that make up this construct. Related to these issues is the type of actual interventions that are employed. As Anastasi (1981) notes, different kinds of interventions will have different effects, consequences, and implications on test performance. It is generally assumed that most children in American culture have had extensive exposure to standardized forms of testing. Indeed, it is often assumed that children experiencing problems in school will have more experience with tests than those not having problems. Jensen (1980) notes:

> Since the 1950s, virtually all children in the public schools have been increasingly exposed to standardized scholastic aptitude and achievement tests, from the primary grades through high school and college, so that exceedingly few pupils by age 10 or so could be regarded as naive in respect to tests. Because of the concern of teachers and parents, the least able pupils or those with special learning problems are apt to be tested the most, especially on individual tests given by a school psychologist. Therefore, it seems most likely that in the present day very little of the variance in standardized aptitude or achievement test scores can be attributed to individual or group differences in test sophistication, with the exception of recent immigrants and persons who have little or no formal schooling or who have gone to

quite atypical schools (p. 591).

Nevertheless, the degree to which children have previous experiences with individual intelligence tests may vary and brief orientation sessions may be effective in equalizing test sophistication. To the degree that preexisting differences can be reduced through test orientation, a more valid measure should be obtained (Anastasi, 1981). Also, as Jensen (1980) notes, coaching and practice may help "equalize test sophistication among persons with differing amounts of past-experience in taking standardized tests or who differ in the recency of tests" (p. 596). Thus, validity may actually be enhanced to the degree that test sophistication differences are minimized or eliminated. However, this should be done when evidence has been gathered that individuals or groups have little or no test sophistication. Methods for doing this are discussed later in the report. Also, in cases of severe problems with orientation to a test, more specific interventions may be necessary (see later discussion).

Another issue related to interventions on tests is that the effect of training may be specific to the skills used during training. In some respects the training conducted on test items is very similar to many intervention programs in behavior therapy or behavior modification where generalization has not always occurred. Indeed, unless specific attempts are made to facilitate generalization, it is likely that it will not occur (Stokes & Baer, 1977;

Wildman & Wildman, 1975).

Aside from the interventions focused on specific test items, the focus of training efforts have also been on more broad-based cognitive skills. In this case virtually any intervention that could be implemented with cognitive skills could be used. These more broad-based cognitive skill training programs are discussed later in the chapter (see pp.000). Suffice it to say that the focus of such programs raises issues of why tests should be used as a dependent variable when the broader curriculum is the real (and correct) focus of improvement.

A final consideration in this area relates to the methodological problem in studies that have been published to date. Studies in this area are far from methodologically pure, although there are some well designed investigations. Future research would need to consider several issues (cf. Anastasi, 1981): Namely, inclusion of a control group, random assignment of subjects to groups, the comparability of pretest and posttest sessions with regard to maturation to perform well, and the assessment of generalizability or transfer of training to nontest performance.

Motivational and Situational Factors

A variety of motivational and situational factors have been discussed within the context of test bias. These include motivational manipulations, text anxiety, modifications in test procedures, and various situational and procedural issues.

Motivational Factors

Numerous studies have been published that focus on motivational factors that may increase test performance. Considerable diversity of procedures, subjects, tests, and incentives characterize these attempts (Kratochwill et al., 1980). Some studies have reported that when test responses are reinforced, performance is higher than previously (and/or the performance of control subjects under standard conditions) (e.g., Ayllon & Kelly, 1972; Bergan, McManis, & Melchert, 1971; Edlund, 1972; Hurlock, 1925). However, the results are not always in favor of reinforcement. For example, some researchers (e.g., Benton, 1936; Maller & Zubin, 1932; Tiber & Kennedy, 1964) found no significant difference in performance between subjects tested under standard conditions and those tested under reinforcement conditions. Also, Clingman and Fowler (1976) studied the effects of candy reinforcement on IQ test scores in first and second graders. No differences were found among these conditions (candy given contingent on correct responses, candy given noncontingently, or no candy given) on test-retest administrations to the Stanford-Binet (Form L-M).

Smeets and Striefel (1975) compared deaf children's scores on the Raven Progressive Matrices (Raven, 1938) when tested under (1) end-of-session reinforcement, (2) noncontingent reinforcement, (3) delayed reinforcement, and (4) immediate contingent reinforcement. The authors found that while the mean posttest score of subjects tested under the immediate-reinforcement condition was significantly higher than that of any other groups, no significant differences were observed among the mean posttest scores of the three other groups.

In the Clingman and Fowler (1976) study, the authors also compared the effects of contingent candy reward, noncontingent candy reward, and no candy on the IQ scores (PPVT, Forms A and B) of children whose initial scores placed them in three different IQ levels. Results showed that candy administrated contingent upon each correct response increased IQ scores for the initially low-scoring subjects, but had no effect on the scores of middle and high-scoring subjects.

Some minority group studies have shown that reinforcement (e.g., praise or candy) did not affect black children's Stanford-Binet scores (Quay, 1971; Tiber & Kennedy, 1964) and that feedback and reward led to significantly higher WISC Verbal Scale scores of lower class white children, but not of lower class black or middle class white children (Sweet, 1969). Cohen (1970) found no significant interaction of verbal praise and candy incentives with whites and blacks on the WISC Block Design performance of second and fifth graders. Also, Wenk, Rozynko, Sarbin, and Robison (1971) found that there were no significant interactions of race with (a) maternal reward, (b) verbal praise and encouragement, and (c) no incentive on white and black delinquent male youths on performance on four of the nonverbal performance tests of the General Aptitude Test Battery.

In contrast to these findings, Klugman (1974) found that in 7 to 14 year-old children a money incentive (as compared to praise) for correct responses on the Stanford-Binet improved the mean IQ of blacks by 4 points, while no significant effects were found for whites.

Considerations

The usual conclusion from this literature is that the effects of incentives, on minority students, is negative (with the exception of Klugman, 1974) (Jensen, 1980), or that the use of incentives does not increase children's scores over and above the usual testing situation which may itself be somewhat motivating (Sattler, 1982). Actually, the issue of whether or not incentives have any effect on test performance is not easily discernable based on the existing literature. Several issues need to be addressed in this area (Kratochwill et al., 1980). To begin with, variations among studies also make trends difficult to identify. For example, some studies have focused on the effects of certain types of reinforcers, such as praise (Bergan et al., 1971; Hurlock, 1925; Roth & McManis, 1972; Tiber & Kennedy, 1964) and candy (Edlund, 1972; Tiber & Kennedy, 1964) on test performance. Vastly different conditions have been developed to represent the "reinforcer."

There have also been variations in procedure. For example, in some studies children received reinforcement immediately after every correct response (e.g., Bergan et al., 1971; Edlund, 1972; Roth & McManis, 1972), while in other studies they received reinforcement after every subtest or when the test was completed (Ayllon & Kelly, 1972; Hurlock, 1925; Tiber & Kennedy, 1964). Yet in other investigations, subjects were promised reinforcement if they performed better (Benton, 1936; Maller & Zubin, 1932). Thus, variations among studies make any clear trend difficult to determine.

Another major concern is that studies have not indicated what reinforcement procedures constitutes an optimal motivational condition. As traditionally conceived, positive reinforcement refers to an increase in the frequency of a response following the presentation of the event (Kazdin, 1980). Whether or not one can identify an event as a positive reinforcer is determined empirically by examining the relationship between the event and behavior. The point is that it is questionable whether or not the vast majority of published studies in this area have adequately tested reinforcement effects on test performance.

Current problems in this area are not likely to be addressed with the exclusive reliance on Large N between-group methodology (Kratochwill & Severson, 1977). Thus, there may be no best reinforcer for a random group of children (Parton & Ross, 1965). Also, Schultz and Sherman (1976) were unable to draw any firm conclusions regarding reinforcement after reviewing approximately 60 studies in the reinforcement literature. Similar to others (e.g., Bisett & Rieber, 1966), they concluded that reinforcers should be individually determined rather than depending on a prior assumption. Thus, each child may have a different reinforcer and this would need to be determined. Several procedures could be used to determine this, including reinforcement hierarchy approaches (Forness, 1973), various self-report schedules (Sulzer-Azaroff & Mayer, 1977), and most importantly, an empirical determination of reinforcing events (e.g., Bijou & Grimm, 1975; Bijou & Peterson, 1971; Lovitt, 1975).

Another issue that needs to be addressed in this area relates to whether or not changes in IQ test scores is a relevant focus of efforts. Conner and Weiss (1974) argued that it is unwarranted to assume that an increase in correct responses is necessarily paralleled by an increase in "cognitive ability." Thus, if the effects of reinforcement in test-taking situations are limited to a motivational function, and if all populations from which samples are drawn demonstrate the same increase in motivation, then administration of reinforcement will shift the distribution of scores upward, resulting in each subject's relative position remaining the same. However, as Clingman and Fowler (1976) note, if further research substantiates the notion that only select populations benefit from reinforcement in pretests, then the use of reinforcement would not increase the motivational level of all subjects, but would selectively enhance the performance of children for whose correct responding is not maintained by other than external reinforcement. If this were the case, the validity of the test would be improved by reducing motivation specific effects. If such changes occur, it would likewise be necessary to determine the effect motivation has on the predictive utility of the test. While it can be argued that the use of such reinforcement results in performance that better reflects what the individual actually knows, it may also result in the test becoming less predictive of some external criteria. This would be the case if such reinforcement influences performance on the criterion to a similar degree. In essence, what one may have is a test that better measures the construct it purports to measure but also a test that does not predict the criterion as well as the test that did not induce optimal performance. If such were the case, the validity of the test would be different for different groups and consequently biased. Needless to say, more research in this area is necessary to sort through the various possibilities.

Finally, the issue of whether or not children should be reinforced depends somewhat on the standardization procedures for the test involved. Deviation from standard procedure changes the meaning of scores (Cronbach, 1960), and may actually invalidate test norms (Braue & Masling, 1959; Sattler, 1974; Strother, 1945). Tests may vary considerably in the way this issue is handled. For example, some test manuals encourage examiners to give approval for incorrect as well as correct responses (i.e., "effort"). While this may increase the total number of responses, it may not affect the number of correct responses (Clingman & Fowler, 1976).

Situational Factors

There is a rather large literature on various situational factors related to test performance. Much of this literature has been reviewed (see Anastasi, 1976, Ch. 2; Jensen, 1980, Ch. 12; Sattler, 1974, Ch. 6, 1982; Sattler & Theye, 1967, Ch. 5). An important issue is whether or not any of these situational factors interact with any cultural groups to produce differences on mental test performance (Jensen, 1980). Here it is assumed that if the child does not perform as well as possible during the testing situation, an inaccurate reflection of classroom performance may occur (Reschly, 1979).

Research has been directed at explicating the situational factors which may allow testing situations to yield a valid assessment of the child's cognitive abilities. Some research involving the use of familiar examiners (Thomas, Hertzig, Dryman, & Fernandez, 1971), positive pretest interactions between examiner and child (Jacobson, Bergen, Bergman, Milhaus, & Grerson, 1971), and testing location (Seitz, Abelson, Levine, & Zigler, 1975) has suggested that a situational/motivational explanation for the poor performance of economically disadvantaged minority children is a definite possibility. Also, some research has demonstrated that disadvantaged children dislike evalution (e.g., Labov, 1970; Zigler, Abelson, & Seitz, 1973) and the testing milieu (e.g., Bee, Streissguth, VanEgenan, Leckie, & Nyman, 1970; Johnson, 1974; Labov, 1970). Both direct (Sacks, 1952; Zigler, et al., 1973) and vicarious (Piersel, Brady, & Kratochwill, 1977) models exposed to an examinee prior to a test administration have facilitated performance. For example, Piersel et al. (1977) found that a pretest vicarious situation in which minority group children watched a seven-minute videotape of a white examiner testing a minority child under positive conditions (e.g., praise) resulted in only 14.3% of the WISC-R scores being 1SD below the mean, whereas 42.8% and 52.4% of the scores were 1SD below the mean under standard and feedback conditions, respectively.

Some investigations have shown that low SES preschool children (black and white) obtain higher scores on the Stanford-Binet (Form L-M) when a test administration procedure allows a maximum number of successes early in the testing experience than under standard administration procedures (Zigler & Butterfield, 1968). Performance was optimized by such procedures as presenting easiest items first and giving easier items from an earlier age level wherever the child failed two successive items. Ali and Costello (1971) found that randomizing the difficulty level of the items on the Peabody Picture Vocabulary Test (PPVT), along with other procedural changes, led to higher scores than did a standard procedure in a group of black preschool children. Unfortunately, these studies do not elucidate the race interactions with changes in testing procedures, since the results have not been analyzed separately by race. As Jensen (1980) notes, such departures from standard administrations only show that the changed procedures can have some effect on the average level of the scores. In a study which varied the testing atmosphere (i.e., relaxed game-like versus formal, evaluative) in an administration of the WISC to 208 white and 208 black junior and senior high school students (Samuel, 1977), no significant main effect of the testing conditions and no significant instructions with race of S_s, E_s, sex of E_s, or E's expectation of S's IQ level was found. Yet research in this area does point to the fact that standardized test practices do not induce optimal performance, a fact that needs to be more fully examined for its potentially biasing effects.

Text Anxiety

A considerable body of literature has developed in the area of test anxiety, but relatively few studies have involved an examination of whether the anxiety construct accounts for cultural or racial groups differences in test performance. Major reviews of the test anxiety literature have appeared (e.g., Anastasi, 1976; Morris & Kratochwill, 1983; I. G. Sarason, 1978; S. B. Sarason, Davidson, Lightball, Waite, & Ruebush, 1960; Sattler, 1974, 1982) and Jensen (1980) has reviewed work in the area of test anxiety and bias.

Research on test anxiety and bias is generally inconclusive and has examined a very narrow version of the anxiety construct. Solkoff (1972) administered the Sarason Test Anxiety Scale to black and white children between the ages of 8 and 11. Results showed no significant race difference, no significant interaction with S's race X E race, and no significant correlations with WISC Full Scale IQ. Jensen (1980) reviews two studies in this area by he and his associates. In the first (Jensen, 1973e), a questionnaire measure of manifest anxiety [The N(neuroticism) scale of the Junior Eysenk Personality Inventory] was given to samples of white, black, and Mexican-American children in grades 4 to 8. He found a significant (but small) group difference on this measure, with the whites obtaining higher anxiety scores. Also, there were no significant correlations with verbal and nonverbal IQ and scholastic achievement tests. In a later study, Jensen and Figueroa (1975) examined the interaction between race and immediate versus delayed recall of aural digit series (digit span is purported to be sensitive to measurement of anxiety). However, in a large sample of white and black school children in grades 2 to 8 no significant interaction was found in digit span scores. Similar to these results, Noble (1969) found no differences in pulse rates of black and white elementary school children immediately before and after being individually tested.

Considerations

The conclusion of reviewers of this literature is that there appears to be no "consistent" or "appreciable" differential effect of anxiety on the test performance of whites and blacks (e.g.,Jensen, 1980). From the available literature, this must be the conclusion. However, we should point to several problematic issues in this area. To begin with, there just isn't enough empirical work to draw any firm conclusions because "anxiety" assessment has been confined to a rather narrow range of measures. For example, Noble (1969) measured pulse rate, but this is only one of several physiological measures that could be employed. Likewise in the Jensen (1973e) and Jensen and Figueroa (1975) studies only a limited measure of the so-called anxiety construct was employed. Thus, investigation in this area suffers from a construct validity problem in as much as it is not at all clear whether or not anxiety was even assessed.

In order for a reasonable assessment of anxiety to occur, measures should be taken on cognitive, behavioral, and physiological dimensions. Assessing these three response modes provides a more adequate test of whether or not anxiety occurs on more than one measure (Morris & Kratochwill, 1983). Also, it is important to assess each of these measures through a device or procedure that measures some aspect of the three construct dimensions. That is, for example, physiological arousal can be measured through either behaviors, self-report, or physiological equipment (e.g.,GSR, heart rate, blood pressure). Until investigations in this area take into account advances in anxiety assessment, little light is likely to be shed on the role of test anxiety in assessment bias.

Other Variables

Achievement Motivation. Achievement motivation (N-Ach) has been identified as a possible source of assessment bias because it is noted that various cultural groups may differ in their level of n-Ach (Chapman & Hill, 1971). The n-Ach construct is said to influence test performance (a) by determining the level of motivation (e.g., interest, effect, etc.) during development and prior to taking various mental tests, and (b) by influencing motivation during the actual test (Jensen, 1980).

Jensen (1980) noted that conclusions on the role of "achievement motivation as a factor in systematic group biases in testing are virtually impossible in terms of the empirical evidence" (p. 616). There are several reasons that have been identified. First of all, there is, as in the case of anxiety, problems in how the construct has been defined and measured. Second, many of the measures that have been used to measure this construct (e.g., projective tests) suffer from reliability and validity problems. Third, many investigations apparently do not show a strong correlation between n-Ach measures and intelligence tests (Heckhausen, 1967). This latter finding has led Jensen (1980) to speculate that high n-Ach is more a product of high ability than the reverse. Thus, no evidence can be advanced that supports the role of n-Ach in assessment bias.

Self-Esteem. Self-esteem is another construct that may contribute to assessment bias (Jensen, 1980), but unfortunately, no empirical work has examined this possibility. If one particular racial or minority group were to have different levels of this construct than a majority or white population, a case for assessment bias could possibly be built. Jensen (1980) reviewed procedures for testing this hypothesis and the reader should consult the review for specific recommendations.

Reflection-Impulsivity. Considerable empirical work has been conducted in this area (see Messer, 1976 for a review). The assumption in work in this area is that some individuals are reflective in response style on certain tests. Characteristically they would ponder alternatives before responding. In contrast, impulsive individuals are quick to respond and may fail to weight all the alternatives. Kagan's Matching Familiar Figures Test (MFFT) is a common measure of this construct. Unfortunately, this is again an area where virtually nothing is known about its influence in assessment bias (Jensen, 1980). Jensen (1980) does speculate that reflectivity is highly related to g.

Race-of-Examiner

Issues

The race of the examiner has often been proposed as one major source of bias in the assessment process. Indeed, one primary procedure that has been suggested in the spirit of meeting nondiscriminatory assessment criteria is to use a minority group examiner to assess the minority child (Kratochwill, et al.,1980). The tactic is not to just use a minority examiner, but rather one that matches or closely approximates the minority status of the child being assessed.

The examiner's race has been hypothesized to be an important factor in affecting the minority child's test performance through (1) the possibility that the child's perception of the testing situation leads to inappropriate behaviors which are judged by the testor to reflect low ability, and (2) the possibility that final scores are biased by the examiner's expectancies for performance of minority children resulting from pretest referral information and unfamilarity with the examiner's cultural background and dialect (Meyers et al., 1974, p. 22). In practice, this concern has been translated into some specific actions. For example, Garcia (1972) noted: "Be skeptical about utilization of standard diagnostic instruments when used to identify the learning behaviors and capabilities of bilinguals. Instead, utilize bilingual clinicians to assist in the identification process" (p. 3). Most recommendations of this sort relate to individual or one-to-one forms of testing or at least involve some form of interpersonal relationship.

Over the years numerous authors have suggested that racial differences may affect the examiner-examinee relationship (e.g., Anastasi, 1958; Anastasi & Foley, 1949; Garth, 1922; Helgard, 1957; Klineberg, 1935, 1944; Pettigrew, 1964; Pressy & Teter, 1919; Riessman, 1962; Strong, 1913). Some authors have noted that ethnic differences can create an "atmosphere bias" and this should be considered a part of the domain of test bias (e.g., Flaugher, 1978). Flaugher (1978) noted that the very art of testing itself may be unfair to certain minority individuals because the situation itself inhibits usual or typical perforamnce. It is certainly possible that any bias in assessment could be reduced if the examiner possessed a language, value system, cultural information, and a familiarity with learning strategies similar to those of the client.

The conclusion that the race of examiner is a potent factor in test bias is not at all clear. There are both methodological and conceptual factors that have a bearing on any conclusions in this area. One of the most careful examinations of empirical research in this area is presented by Jensen (1980). From research conducted between 1930 and 1977, Jensen classified studies into three main categories of experimental design: (1) inadequate designs, (2) adequate but incomplete designs, and (3) adequate and complete designs. Jensen (1980) defined an adequate design in the following way:

> To be an adequate design, an experiment on the effect of the race of examiner (RE) on test scores should meet the following two minimum requirements: (1) at least two (or more) Es of each race and (2) random assignment of subjects (Ss) to Es. These requirements seem obvious. If there is only one E of each race, the variable of race is wholly confounded with the other personal attributes of each E. Randomization is needed to rule out the possibility of any selection bias that might result in a spurious (i.e., noncausal) correlation between Es and the part being measured. Any study that does not meet these minimal requirements of experimental design is classified as inadequate. When it is not clear whether the study meets these requirements, I have given it the benefit of the doubt and classified it as adequate (pp. 596-597).

Jensen (1980) also determined that an adequate design is _incomplete_ in the case where subjects are sampled from only one racial group and _complete_ when _S_s are sampled from two or more racial groups. A good design (nonrepeated measures) is presented in Table 5.1. In this design it is the interaction

Table 5.1

Adequate and Complete Design for Assessment of Race of Examiner Effects

		Race of Examiners			
		Majority		Minority	
		E_1	E_2	E_3	E_4
Race of Subjects	Majority				
	Minority				

Source: Adapted from Jensen, A. R. *Bias in mental testing*. New York: Free Press, 1980.

of S's race and E's race that is critical to testing the effect of interest.

Based on these organizational formats, Jensen (1980) has classified existing studies (see Table 5.2) and makes various conclusions from investigations within these areas. In the area of inadequate designs, virtually no conclusions can be made, although about half support the hypothesis that black Ss perform better when tested by a black E than when tested by a white E.

An adequate, but incomplete, design format indicates no significant effect of race of the E or S's test performance and one shows a significant effect of race of E. In a more recent study that fits into this conceptualization Terrell, Terrell, and Taylor (1980) investigated the effects of race of examiner and type of reinforcement on the intelligence test performance on lower-class black children. The authors found that children given tangible rewards, regardless of race of examiner, obtained significantly higher scores than did children given no reinforcement or children given traditional social reinforcement. Moreover, the children given culturally relevant social reinforcement by a black examiner obtained significantly higher WISC-R test scores than did children given culturally relevant reinforcement by

Table 5.2

Race of Examiner Investigations

Inadequate Designs	Adequate but Incomplete Designs	Adequate and Complete Designs
Canady (1936)	Smith & May (1967)	Miller & Phillips (1966)
Passamanick & Knoblock	Pelosi (1968)	Abramson (1969)
Forrester & Klaus (1964)	Costello (1970)	Crown (1970)
LaCrosse (1964)	Dell (1972)	Dyer (1970)
Pettigrew (1964)	Moore & Redish (1974)	Gould & Klein (1971)
Lipsitz (1969)		Veroff, McClelland, & Marquis (1971a, 1971b)
Caldwell & Knight (1970)		Yando, Zigler, & Gates (1971)
Scott, Hartson, & Cunningham (1976)		Solkoff (1972)
		Savage & Bowers (1972)
		France (1973)
		Wellborn, Reid, & Reichard (1973)
		Jensen (1974c)
		Marwit & Neumann (1974)
		Solkoff (1974)
		Ratusnik & Koenigsknecht (1977)
		Samuel (1977)

Source: Adapted from Jensen, A.R. Bias in mental testing. New York: Free Press, 1980.

the white examiner. Thus, in this study, no significant main effect for race of examiner was found.

Finally, in a summary of adequate and complete designs, Jensen (1980) notes that 10 out of 16 show no effect of the race of E X race of S. Studies that demonstrate a significant effect have negligible effects of race of E. Moreover the race of E X race of S interactions reduces the overall mean difference between racial groups, a finding in three studies.

Considerations

Based upon work in this area, Jensen (1980) concluded that there is no support for the assumptions that the race of examiner is an important source of variance between whites and blacks on measures mental ability (p. 602-603). While such a conclusion seems possible at this stage of knowledge, several conceptual issues must be raised in this literature. To begin with, there are problems with attempts to compare different studies using a "box score" approach (Kazdin & Wilson, 1978). It is not possible to review all the difficulties with the box score approch here and the interested reader should consult Kazdin and Wilson (1978) for an excellent discussion of these problems in the psychotherapy literature. However, some points might be raised. [Actually,many of the criticisms of the box score approach also apply to the meta analyses (cf. Smith & Glass, 1977) alternatives that are proposed for literature reviews as well.] First, a series of studies that are methodologically inadequate is no better than one study that is inadequate. Jensen (1980) went one important step beyond this and made important distinctions between adequate and inadequate studies as well as between studies that are complete and incomplete. However, another problem emerges with use of the box score approach, even when distinctions are made between "good" and "bad" studies. Even within given studies in these three categories, variations in independent variables, experimenters, children, and dependent measures (i.e., measures of cognitive ability) render any firm conclusions difficult.

A third major problem with the box score approach is that various independent variables are probably not the same. Thus, "race of examiner" and "race of child" are perceived as homogeneous. But, minority group children represent a heterogeneous population, so it becomes difficult to know what group or groups of children should be so classified (Sattler, 1974). For example, Valentine (1971) found 14 different Afro-American subgroups in one urban community and each had more or less distinct cultures. Sattler (1974) also noted that the label "minority group children" is typically used to designate individuals whose values, customs, patterns of thought, language, and interests are different from the dominant culture in which they live (Liddle, 1967). Included within such a conceptualization would be groups including blacks, Mexicans, Indians, Puerto Ricans, and various subculture whites (Appalachians, foreign born, unskilled laborers, and so forth). The point is that observed effects may reflect other differences such as social class to a greater extent than racial or ethnic differences.

Meyers et al. (1974) noted that "race of examiner" may not in and of itself lead to deviant responses by a minority child. Negative responses may be elicited when certain modes of interaction are initiated (e.g., outright expression of disappointment). They note that personal examining style and the milieu created by a white examiner may be more related to testing behavior than examiner race per se (e.g., Bucky & Banta, 1972; Yando, Zigler, & Gates, 1971).

Finally, it is possible that a box score strategy obscures certain developmental patterns that may operate in this area of research. For example, Epps (1974) noted that data from various studies if this area indicate that the age of the examinee may mediate the race of examiner effect. Thus, it is possible that any negative report of examiners of a different race on black and on white children is strongest in the early years. However, in later years the negative input may decrease and the difference can have a facilitating effect (cf. Katz, Atchison, Epps, & Roberts, 1972). In testing situations where no whites are present, the belief that they are competing with whites rather than with other blacks may have an effect on a black student's performance (Epps, Katz, Perry, & Runyou, 1971). It is possible that with black examiners, the implied comparison may enhance performance. It is also possible that the nature of the effect of implied white comparison is mediated by the subject's perception of the probability that (s)he will be successful in the testing situation. It is also noted that when the probability of success is relatively high, white examiners have a facilitating effect; when the probability of success is relatively low, black examiners may have a facilitating effect. While some studies may support this (e.g., Savage & Bowers, 1972; Watson, 1972), Epps (1974) notes that this area needs to be clarified. Also, the relation between the task itself and the race of examiner and race of comparison effects should be further clarified in empirical research.

Sex of Examiner

Issues

Several reviews have focused on the sex of the examiner and its possible influence in intellectual assessment (e.g., Jensen, 1980; Rumenik, Capasso, & Hendrick, 1977; Sattler, 1974). The general concensus from this literature is that there are no consistent effects of the sex of E. However, Jensen (1980) concluded that some evidence suggests that female Es tend to elicit higher performance than male Es from both males and females.

Epps (1974) has further noted that there is really little known about how the sex of E affects the performance of children or how the E's sex interacts with the S's sex in multiracial or multisocial settings. Research may be limited because either investigations have involved only male examiners or data for both sexes have not been analyzed or reported separately. Jensen (1980) speculated that work finding sex of examiner effects may even be Type I errors.

Language

Issues

Language and related factors (e.g., dialect) have often been examined as they relate to possible bias or discriminating effects in assessment (Jensen, 1980; Kratochwill, et al., 1980). Language was considered an important assessment issue as early as 1910, when large numbers of immigrants came into the United States. In order to make assessment less discriminatory or biased, tests or test directions have been translated into the "primary" or "dominant" language of the client. Several tests (e.g., WISC, Wechsler, 1949; Illinois Test of Psycholinguistic Abilities, Kirk, McCarthy, & Kirk, 1971) have been translated into another language (e.g., Spanish), but the number of such translations as used in the U. S. is relatively small. Nevertheless, the tactic of translating tests with the presumed primary language of the client is one criterion for nondiscriminatory assessment in PL 94-142.

Based on considerations of what effect the language of the examiner or of the test itself may have on the performance of children from a bilingual or non-English background, Jensen (1980 pp. 605-606) drew the following conclusions:

1. The language of test and examiner makes a difference in that various racial and cultural groups (e.g., Mexican-American, Puerto Rican, Native American, Chinese, Japanese) generally obtain higher scores or nonverbal and performance tests than an English verbal test, oral or written.

2. On English language tests of scholastic achievement, students with foreign language backgrounds usually perform better on arithmetic than on language items.

3. Generally, Mexican-American individuals score higher on the Wechsler and Stanford-Binet IQ tests when these are administered in Spanish rather than in English.

4. Generally, the language spoken by the examiner makes less difference on performance on nonlanguage tests than on verbal tests (oral or written).

5. When Spanish, Mexican-Americans are equated with Anglo whites and Orientals on socioeconomic status, the lower performance of the former is greatly reduced.

6. Mexican-American children from bilingual homes where both Spanish and English are spoken typically perform better on various standardized tests than children from homes where Spanish is spoken exclusively.

7. Overall, the language of the test or examiner makes less of a differnece on performance the longer the child has attended English language schools. Also, the differences usually found between verbal and nonverbal tests declines

with the increasing number of years in school.

The assessment of bilingual children has special implications for the area of nonbiased or nondiscriminatory assessment. Jensen (1980) noted that while verbal IQ tests given in English to bilingual children is a relatively good short term predictor, such tests should not be used for predictions for special classes when more than one year placement would be made.

In order to be sensitive to bilingual students, a common strategy is to translate the test or administer it in more than one language. However, there are several difficulties that may emerge when this alternative is pursued (Kratochwill et al., 1980). To begin with, the examiner must first determine the primary or dominant language of the child. This is not straight-forward. The lack of adequate language assessment instruments has often hindered assessment efforts as well as the implementation of special language programs and identification of eligible students. The major problems include (1) determination of what language skills and linguistic structures to describe and (2) the identification of adequate tools or instruments to measure language (Silverman, Boa, & Russell, 1976). These authors published the Oral Language Tests for Bilingual Students in an effort to address the policy advanced in the Bilingual Education Act of 1974. Silverman et al. (1976) evaluated various language assessment devices on dimensions of validity, technical excellence, and administrative useability. The evaluation was conducted on commercially available tests, tests under development or undergoing field testing, and tests used for experimental purposes.

Results of this evaluation suggested that of the 24 tests reviewed, only a very few could be used for languages other than Spanish (e.g., MAT-SEA-CAL Oral Proficiency Tests, 1976). Another problem was that the tests reviewed had a restricted age/grade range.

The concept of bilingualism also presents other difficulties in a practical area. Some children may use English in school and Spanish outside school (home and community). Such children may fail to develop a sufficient mastery of either language (Sattler, 1974). For example, in some studies in this area, Spanish has been used either in test directions only or in the complete test to administer standardized intelligence tests to Spanish-speaking children (e.g., Chandler & Plakos, 1969; Galvan, 1967; Holland, 1960; Keston & Jimenez, 1954). After reviewing these studies, Sattler (1974) suggested that such procedures are not only frought with hazards, but that "translations of a test makes it a hybrid belonging to neither culture" (p. 39). Furthermore, whether or not bilingualism will constitute a problem for the child will depend upon the way the two languages are acquired (Anastasi & Cordova, 1953). Sattler (1974) also argued that a child who learns two different languages (i.e., one at home and another at school) may have more problems than the child who learns one language that is expressed across all situations.

While the issues surrounding bilingualism are less than straight forward, the problems are even greater in polylingual cases. For example, polylingualism may complicate language assessment and could even be related to observed speech difficulties (Sattler, 1974). Specifically, it is possible that various patterns of speech developed in the use of one language can interfere with correctly speaking another (Bebevfall, 1958; Chavez, 1956; Perales, 1965). Children may never become proficient in speaking either language (Holland, 1960), and in the case of Spanish-speaking groups, children may borrow from a limited English vocabulary to complete expressions begun in Spanish. They may give English words Spanish pronunciations and meanings and they may have difficulties in pronunciation and enunciation (Perales, 1965).

In summary, translations of a test may provide a promising alternative to reduce bias in the assessment process. However, mere translation of the test into the "primary" language of the child has several conceptual and methodological problems that has not been adequately addressed in research in this area.

Dialect

Some minority groups speak English, but there is a clear dialect difference from standard English. For example, many black children speak a form of black dialect English that varies considerably from the standard English spoken by many white children. Oakland and Matuszek (1977) noted that language biases may be encountered in assessing black children who manifest elements of nonstandard dialects. Presumably some tests use language styles that are significantly different from those manifested by blacks (and other minorities), in which language patterns also are ordered and rule governed (Bartell, Grill, & Bergen, 1973; Gay & Abrahams, 1973). Dialect differences are not limited to racial groups. Many whites from certain parts of the country or various SES levels speak with a dialect that varies from that spoken in the majority white culture. The issue, no matter what the dialect, is whether or not differences on this dimension influence performance on standardized tests in a way that will bias decisions.

A point has been made that even if English is the primary language, there is considerable variations among cultural groups in terms of complex language idioms, colloquialisms, words and phrases with multiple meanings, and words and phrases of similar but not identical meaning within a language (Garcia, 1976). It has also been verified that even if English is the primary language, testing procedures may not equate for differing cultural or subcultural information learning strategies, and value systems (Alley & Foster, 1978).

Nevertheless, Jensen (1980) noted that a number of studies have suggested that black children comprehend standard English at least as well as their own nonstandard dialect and that their understanding of standard English occurs at an early age (e.g., Eisenberg, Berlin, Dill, & Sheldon, 1968; Hall & Turner, 1971, 1974; Harmus, 1961; Kraus & Rotter 1968; Peisach, 1965; Weener, 1969).

While many important conceptual issues have been raised in this area, the empirical literature provides no support for the effect of dialect (Jensen, 1980). In an early study, Crown (1970) studied the effect of language dialect (black versus standard English) and race of examiner (two black and two white examiners) on the Wechsler Preschool and Primary Scale for Intelligence (WPPSI). The results showed no effect of dialect and no interaction with race of examiner or race of student. Similar results that do not support dialect effect have been found in a series of studies by Quay (1971, 1972, 1974) on the Stanford-Binet Form L-M. Thus, results of empirical work in this area do not support the notion that dialect influences test performance. However, there is relatively little work in this area.

Bias in Test Scoring

Scoring bias refers to any systematic error that occurs in deriving the scores from the test (i.e., systematic errors in scoring). Research in this area has usually found some halo effects on such tests as the Stanford Binet and Wechsler scales. For example, in the usual procedure in this area, examiners are given expectancies that a child is bright or dull with various ambiguities in various response items. In some studies where expectations are manipulated, examiners tend to overrate ambiguous responses for high expectancy subjects and underrate them for low expectancy subjects (e.g., Egeland, 1969; Sattler & Winget, 1970; Sattler, Hillex, & Neher, 1970; Simon, 1969).

There are at least three problems in work in this area (Jensen, 1980). First of all, research has generally produced effects that are of little or low magnitude. Second, and perhaps more important, the research has usually been conducted under more analogue conditions. The expectancies are contrived and the study is conducted under laboratory or non-field conditions. Thus, it is not at all clear that the results would occur under conditions present where IQ tests are usually administered (e.g., school settings). Finally, studies demonstrating halo effects have usually failed to determine if test validity is compromised. Jensen (1980) notes:

> The most telling experimental paradigm, which has never been applied, would be to substitute a small number of ambiguous responses made in authentic test protocols ranging widely in total score and note the degree of discrepancy between ratings given to the substituted ambiguous responses and the ratings given to the S's actual responses on these items. Based on probability, it is likely that the halo effect, on the average, enhances the scoring validity of highly ambiguous

responses (p. 610).

Some research has also been conducted on the effect of scoring bias as a function of race (e.g., Jacobs & DeGraaf, 1972). In this study, there was a main effect for expectancy (high expectancy X IQ = 89.6 and low expectancy X IQ = 87.9). There was no significant interaction with race of subject or race of examiner, and no interactions among these factors.

There is also some evidence to suggest that examiners give higher estimates of intelligence for blacks and Mexican-American children than for white children with the same measured IQ (Nalven, Hofmann, & Bierberger, 1969; Sattler & Kunck, 1976). Jensen (1980) noted that such results may indicate that some psychologists either accept the notion that tests underestimate the IQ of minorities or that more weight is given to various ability factors.

Bias in Observational Assessment

Bias in assessment is not limited to standardized ability measures as usually conceived. Such assessment procedures as direct observations in naturalistic settings have also been examined for bias. Indeed, a number of authors have provided reviews of this literature discussing such factors as interobserver agreement, training observers, code complexity, and communication among observers (e.g., Johnson & Bolstad, 1973; Kazdin, 1977; Kent & Foster, 1977; Wasik & Loven, 1980; Foster & Cone, 1980; Wildman & Erickson, 1977; Haynes & Wilson, 1979).

One strategy to investigate bias in observational assessment is to create expectancy instructions regarding changes in the clients being observed. In an early study in this area, Kass and O'Leary (1970) informed observers that the frequency of descriptive behavior of children should either increase (group 1) or decrease (group 2) during treatment. A third group was told that the researchers were unsure of the effects of the treatment. The authors found that all groups recorded a decrease in disruptive behavior with the treatment, with the largest effect occurring in the groups of observers who were provided the expectation that the frequency of the disruptive behavior would decrease. Although it is possible that this effect exists, this study has been criticized on methodological grounds (e.g., Johnson & Bolstad, 1973; Kent, O'Leary, Diamont & Dietz, 1974) and has not been replicated in two attempts (Kent et al., 1974; Skindrud, 1972 .

Bias in observational assessment has also been studied by providing observers differential feedback concerning their conformity to ratings provided by the experimenter (O'Leary, Kent, & Kantowitz, 1975). In the study, observers were told that a decrease was expected in the frequency of occurrence of two categories of behavior. They were also informed that no change was expected in the frequency of the other categories. The authors found significant decreases in the frequency of the categories for which a decrease in frequency was predicted. Also, no differences were found for the two other categories. Thus, differential feedback and prediction of a decrease led to biased assessment. However, a control group which received no feedback should have been included to examine the insufficiency of predictions alone (Wildman & Erickson, 1977).

This sampling of work in this area conveys something of the flavor of factors that may bias observational assessment. Specific recommendations for obtaining more valid and reliable data through observational procedures are presented in Chapter 7 (pp. 000-000).

Bias Due to Timed vs Untimed Testing

Whether or not a test is timed or untimed (speed vs power tests) has sometimes been postulated as a factor contributing to bias in testing. However, empirical work in this area has not supported this possible source of bias. For example, Dubin, Osborn, and Winich (1969) studied the effect of time limits on the testing of black and white high school students of low and high SES groups . The authors found that both whites and blacks obtained higher scores as a function of practice and an extended time limit. Thus, the findings indicate that black subjects (and low SES) were not penalized when given no extra practice for speeded tests.

Although there is no evidence for bias through time factors in tests, there is very little work in the area. Jensen (1980) noted that two types of speed factors have been identified following Spearman's (1927) work in this area. One, "speed of cognition" refers to the speed with which an individual recalls relevant information for answering a question or for solving a problem and the speed of mentally manipulating the elements of the problem in working toward its solution" (pp. 611-612). Spearman also noted that individuals could have a preference for speed in performing a certain task, a speed factor Jensen (1980) has labeled "personal tempo". However, Jensen (1980) noted that no evidence supports the notion that a personal tempo factor (in contrast to cognitive speed) contributes to any meaningful difference between test scores of various racial and socioeconomic groups.

Summary and Conclusions

In this chapter we provided an overview of situational bias in assessment. As noted in the introduction to this chapter, situational bias and assessment refers to those features that are a part of the assessment process but have not been specifically considered in terms of the technical test bias features described in the latter part of the report. In the chapter we provided an overview of test sophistication. This area included practice, coaching, test sophistication, and interaction with race and social class. We noted that generally there is a paucity of information to suggest that these features bias tests in any systematic way. Nevertheless, there is need for future research in this area.

A host of motivational and situational factors in the assessment process were reviewed in the chapter. These included such things as motivational components, (e.g., reinforcement and incentives), situational factors, test anxiety, and a number of other variables including achievement, motivation, self-esteem, reflectivity, and impusivity. In this area we were impressed with the lack of empirical information pointing to any strong influence in motivational and situational factors. Nevertheless, we must emphasize that the fact that studies are not supportive of one particular direction, does not necessarily mean that these factors can be eliminated as potential candidates for assessment bias. Indeed, in some areas such as the reinforcement literature, adequate tests of the motivational components have really never been tested due to problems in the way studies have been conceptualized. It appears that an individual analysis of motivational factors could likely yield important data regarding the influence of these factors in test bias. Nevertheless, we must again point to the need for future research to further elucidate different motivational and situational factors in the test bias literature.

In the next area race of examiner was discussed. We conclude, as have other researchers in this area, that there is insufficient data at this point to draw firm conclusions regarding a race of examiner effect. Similar conclusions can be drawn in terms of sex of examiner issues.

Another area where situational bias has been examined is in language considerations. Language has been explored in more detail than some of the other areas, but again, there are very few studies that indicate that language is a sole biasing feature when other variables are considered. However, at the most straightforward level, administering a test in English to a child whose language is other than English certainly could be considered bias in assessment. Yet, when some language factors are considered, the role of language factors in assessment bias becomes even more complex. We pointed to some areas of future researh in this area hoping that some new areas of investigation could be opened.

Finally, several other areas of potential bias in assessment were discussed, including bias in test scoring, observational assessment, and potential bias due to timed vs. untimed testing. Work in each of these areas is relatively primative at this time. However, at this point there is no clear evidence that these factors have resulted in systematic situational bias in assessment.

After reviewing the rather extensive literature in this area, we have to conclude that it is not a matter of more research in each area, but rather the specific type of research that needs to be conducted in the future. Various areas for future research were outlined.

Chapter 6

Outcome Bias

The concept of validity as traditionally conceived focuses almost exclusively on how well a test measures the construct or latent trait it purports to measure. The need for such validation work is obvious. The goal in test development, albeit never reached, is to create a test that is perfectly correlated with the construct it measures. Validation efforts, traditionally conceived, focus exclusively on demonstrating the correlation between the two. With respect to bias, the research efforts discussed in Chapters 4 and 5 have examined the validity of tests to determine if they are measuring the same construct across groups and if so, are they equally valid for all.

From an alternative perspective, that concept of validity which focuses on demonstrating the correlation between the test and the construct it purports to measure can be viewed as parocial. It can and has been argued that the concept of validity should be broadened (Cole, 1981; Cronbach, 1980; Messick, 1975). While studies conceived within the parocial concept of validity provides us information to help explain why an individual performs the way he/she does on a test (i.e. he/she possesses the construct to a certain degree), they tell us practically nothing about the valid use of the test. When tests are used for decision making that results in the selection of individuals, the planning of treatment for individuals, or both, it is vital that information be available that provides the decision-maker(s) the most valid information on which to base a decision, that is, information that predicts the desired outcomes with the least amount of inference. When we add the notion of test use to our concept of validity, by necessity, we focus on whether or not the outcomes of the process employing the test are desired. In order to collect data on the validity of a test under this broadened concept of validity, one would have to know precisely the desired outcomes, that is, the purpose for the test's use. Consequently, under the broadened conceptualization, there is no such thing as a valid test; only tests that are to some degree valid for a purpose. Likewise, a test can have both validity and be invalid according to the decision one makes with it and the desirability of the outcomes of those decisions. In addition, under the broadened concept of validity, when we want to study bias we are interested in whether or not the outcomes are the desired outcomes for all groups.

When traditional validity approaches use external criteria against which to validate tests, the effort can be viewed as an attempt to demonstrate the test's relationship to criteria in which the construct is hypothesized to be associated. Predictive validity studies that employ external criteria do have a practical side to them since the external criteria employed is often an important criteria to which one predicts in decision-making. For example, validating an intelligence test, in part, by demonstrating its relationship to a standardized academic achievement test not only shows the test is acting the way it should, given the construct is measuring, but it also provides information on the relationship of the test to an important criteria (e.g., academic achievement) such information is useful when making decisions about special education placement. Information on this relationship helps us increase the probability of making a correct decision.

Yet, when we take a closer look at a typical decision of this sort and focus on the intended outcomes, we see the large inferences in the interpretation of test data when we rely solely on information provided by predictive validity studies. Three major areas in which we lack information can be identified.

Predicting Specific Outcomes

The first area where there is a paucity of information involves how much is known about the relationship between the desired outcomes and the test. For inferences presently made with tests to be reduced in size, one needs to clearly identify all desired outcomes and have empirical information on the relationship of the test to a criterion that best defines the outcomes. In our example above, if one desired outcomes is to provide an effective intervention through educational placement, then one needs to gather information on how well the test predicts the effectiveness of the placement (i.e., the desired outcome) with effectiveness embodied and defined in a criterion or set of criteria. It is evident that information from predictive validity studies, traditionally conceived, provide minimal information on the use of tests for this purpose. Predictive validity information does provide the advantage of making predictive statements regarding how well the child will perform in the future on standardized tests of academic achievement but provides no information on the test as it relates to the effectiveness of the placement.

With respect to the latter, predictive validity information tells us how well the child will perform without placement, it does not tell us how well the child will perform with placement. If the desired outcome of a placement decision is to help the child learn more effectively, then being able to predict this from a test is of much importance to the decision maker and within the purview of a concept of validity, broadly defined.

In addition to the above problem in predicting to one criteria(e.g., standardized achievement tests) across different conditions (i.e., with and without placement) is the problem of the criteria to which one predicts. Certainly, when deciding on special education placement, standardized achievement tests are only one measure of academic performance that can be used. Others such as work samples, actual learning, and teacher ratings may also be important to use as a criterion measure. This is a question of the validity of the criteria used and such validity can only be determined after a thoughtful analysis of the purpose of assessment. Once a criterion or set of criteria are decided upon, the validity of the predictor measure (e.g., intelligence test) needs to be validated across placement and nonplacement situations to judge the efficacy of the predictor. Note that in traditional predictive validity studies the criterion is customarily chosen in keeping with the purpose of the validation effort, that is, to show the test predicts the criteria it is hypothesized to predict, not necessarily in accordance with the purpose of any decision to be made.

When ones notion of validity includes outcomes, it may broaden the definition and consequent search for bias in assessment. The concern for bias under such circumstances would entail whether or not the test is valuable in predicting equally well the effectiveness of placement across groups. Once an appropriate criterion or set of criteria have been identified, a test used to predict the effectiveness of placement should be demonstrated to be equally effective for both minority and nonminority children.

Tests Within the Assessment Process

The second area in which there is a lack information involves the various forms of data that are considered in decision making; that is, understanding of the assessment process. As conventionally defined, an assessment process is that process of collecting data for decision making (Cancelli & Duley, in press). Discussions of validity to this point have focused on the validity of tests. However, tests are but one element of the assessment process. A wealth of data are usually employed to predict outcomes and to make decisions. In our example above, the decision to place a child in a special education class involves, by law, the efforts of a multidisciplinary team. Each member brings with him/her relevant (and unfortunately irrelevant) data for predicting outcomes. In addition, some of the data are often subjective and highly situation specific. Data can include the subjective impressions of the personality characteristics of the special education teacher, the student, and the interaction of the two, the characteristics of the students in the class with whom the child may be placed, the cooperation of the parents with the placement and so forth.

How all these data fit together within the dynamic process of teaming in making a prediction about the success of placement is, of course, a validity question when one includes in the concept of validity evidence of the effectiveness of decisions as judged by desired outcomes.

Other Considerations Within the Decision-Making Process

The third area in which we have little information involves how considerations other than those drawn from psychological and educational data are brought to bear on the decisions that are made. These considerations often are independent of the predictions of the effectiveness of outcomes as determined through psychological measures. Some of these issues evolve out of ethical, moral,and legal standards and the value of their use is judged in terms of social value (Messick, 1975). The societal impact of placing a disproportionate number of minority students in classes for the mentally handicapped is an example of one such consideration.

Other such considerations evolve out of practical features of assessment. Whether or not a more valid/less cost efficient or less valid/more cost efficient assessment battery should be used is one example of a practical consideration. Still other considerations stem from our concern for the integrity of the decision making process. These issues often involve the decision maker(s)s' concern for the less than perfect reliability and validity of the predictors employed and their yet unidentified biased nature. Such concerns also bear the potential of impacting on decisions.

Decisions regarding whether or not to consider these factors has to evolve out of a clear understanding of the purpose and desired outcomes of assessment. Their use can not be determined by scientific inquiry. They are value judgments based on such concepts as equality and fairness. As such, they are not validity questions in a traditional sense. Yet, in the broadest sense such considerations do impact on whether or not certain desired outcomes come about (e.g., proportional representation) and in that sense are questions of validity. Thus, they are just as important, if not more important, to understand than are data collected from tests and other aspects of the assessment process discussed above. The importance of understanding such considerations lies in the amoral nature of psychological assessment data. For example, a technically unbiased test does not guarantee that its use will not result in socially undesirable consequences. The question as to what is socially desirable although based on values should be a concern to all those involved in the assessment process.

As one can readily see, these considerations could be employed to have an impact on the desired outcomes of decisions as they relate to members of various minority groups. As identified above, for example, decision makers may wish to have as an outcome, proportional representation of minority and nonminority children in classes for the mentally handicapped. Such a consideration would have nothing to do with improving predictions that are made regarding the effectivenss of placement but nonetheless may impact on the decision not to place certain minority students in such classes.

Selection versus Intervention

When we turn attention in assessment to the study of outcomes, the type decisions made with the data need to be precisely identified. While the validity of any test or assessment process must be judged within the context of the decision to be made, there are certain type decisions that can be identified for study. Two such type decisions are those of selection and intervention.

The major difference between these two type decisions is in their purpose. Selection decisions are those that require the decision maker(s) to choose whether or not the individual assessed should be selected. Common type selections that employ psychological data in the process are for the purpose of employment and admissions. Both involve decisions to include or not to include and the effectiveness of the decision and antecedent assessment process are determined by whether or not those who are selected are those who the decision makers want to select. For example, in the area of employment testing the effectiveness of a decision to select someone for a job needs to be judged against what the desired outcomes of that decision are. If an employer chooses to select only those applicants who have the best chance of succeeding on the job, then effectiveness needs to be judged by how well the battery of predictors accomplish that goal. If another employer wishes to choose those most likely to succeed within certain defined racial/ethnic groups so as to maintain proportional representation among these various groups, then the effectivenss of the decisions and assessment process needs to be judged against that desired outcome.

Intervention decisions, on the other hand, have a different purpose and therefore different desired outcomes. The purpose of intervention decisions is to provide successful forms of treatment as a consequence of the decision. The type information desired for interventions decisions is that which will enable the decision makers to better choose an intervention and consequently better predict the desired outcome (i.e., effective intervention). All intervention decisions are preceeded by selection. Although selection decisions are a prerequisite to intervention decisions, it is important to maintain the distinction for two reasons. First, the information necessary to validate the decisions are different. When making decisions regarding intervention it is necessary to know if the desired outcome of the intervention can be predicted from the test. In selection decisions it is only necessary to use tests that will predict who will or won't succeed without intervention. Second, the distinction is crucial since different tests can be valid for making different decisions. A test that is valid for predicting future performance on some criterion may be of no value in predicting success of an intervention designed from its use. This point is clarified by closely examining the selection and intervention decisions involved in making an educational placement in a class for the mentally handicapped.

Somewhat different from employment selection decisions which focus on a test or assessment battery able to predict future performance, the selection process in special education involves the use of diagnostic constructs. Not only do we have to predict future performance on a criteria, but we also have to explain why the individual performed in certain ways. Assessment data employed to make decisions for the selection phase of the placement process must therefore be able to predict who will succeed and who will not succceed, and, in addition, tell why those selected didn't succeed. Consequently because of the diagnostic requirements, tests must demonstrate good construct validity as well as validity in predicting criteria relevant to the decision maker.

It is at this point that the influence of considerations other than those derived from test data impact on the decisions to bring about additional desired outcomes, such as proportional representation. Once the decision is made to select, then an intervention needs to be decided on. In our present example this decision usually involves placement in a class for the mentally handicapped. In addition, by law, the intervention decision must be more specific than just placement and include the specification of objectives and instructional strategies to reach the objectives. The assessment data employed to make such intervention decisions must give the decision-maker the ability to predict the success of the intervention. This test is usually different from that used for the selection decision. For example, a intelligence test provide the decision-maker with the ability to predict to a moderate degree future academic success without intervention; and tests also help provide an explanation for why success is or isn't predicted. However, there is no empirical evidence to indicate that employing data derived from an intelligence test is of value in either predicting the success of placement in an FMR class or designing specific intervention strategies. Thus, from a broadened conception of validity, under such circumstances, an intelligence test is both valid and invalid; valid in selection and invalid for making intervention decisions.

The remainder of this chapter focuses on that literature which addresses bias in outcomes for each of the two classes of decisions, selection and intervention.

Selection Bias

The major contribution to the area of selection bias comes from those who have studied the various models that can be used in selection that take into account the social value considerations that we spoke of above. This literature has mostly addressed decisions in the employment and admissions area and has been discussed under the headings of fairness in selection and bias in selection.

Since the models discussed in this literature eminate from considerations of the abstract concept of fairness as it relates to the selection of minority and nonminority applicants, there is no research available to tell us which model is better than another. There is no model that is more fair or less fair than another. With respect to the concept of validity, the various models can either add validity to the decision or not according to what the desired outcomes of the decisions are. Similarly, with respect to bias, the models are nonbiased if they result in the desired outcomes with regard to group membership. The choice of the desired outcome is a value judgment and an issue of fairness; the issue of whether or not the model yields the outcome is within the purview of a broadened concept of validity and is an issue of bias.

Since all the models are based on various notions of fairness, their evolution is based on various philosophies of fairness. Three philosophies of fairness in selection important to our present discussions have been identified by Hunter and Schmidt (1976): Unqualified individualism, qualified individualism, and quotas.

Unqualified Individualism. This philosophy maintains that any predictor variable or set of predictor variables, regardless of the nature of the variable, should be used to predict a criterion if it improves prediction. Predictors may include such data as test scores, demographic information regarding the individual's race, religion or socioeconomic status, and biological information including sex and handicapping conditions. The one stipulation is that the information used must increase prediction of the desired outcomes and desired outcomes are only those that relate to performance on a criterion. So for example, an employer adopting such a philosophy would design an employment battery to collect data on all those variables that are going to help choose those job applicants that have the best chance of succeeding on the criterion or criteria that define success on the job. The practical utility of including predictors that add minimally to the prediction are left to the discretion of the employer. Group membership only comes into play if such membership helps improve the prediction for that individual. If predictors function differently in their relation to the criterion or criteria across various groups, then the best regression equation for each group is used. Individuals who score on the predictor(s) at a level that would allow prediction at a minimally acceptable level on the criterion are accepted, while those who don't are not. If only a few applicants can be selected, then the applicants are selected from the top down from a list ranking the applicants in terms of their predicted performance without regard to group membership.

There are several advantages to such a philosophy. The most obvious advantage is that it guarantees the selection of only those who have the best predicted chance of succeeding on the criterion. Group membership does not enter into the decision once the applicants are ranked, thus avoiding a situation where one applicant is chosen over another soley because of group membership and not merit (as defined by performance on the criterion or criteria). It is argued by those who advocate such a philosophy that this gives members of all groups an equal chance of success that would not be predicted if special advantage was given to one group over another for nonmeritorious reasons. In addition, such a philosophy tends to reduce the mean differences in performance across groups among those selected, thereby more closely approximating equality in future chances of promotion or graduation across groups, and reducing group differences in failure or frustration.

The main objection to those who find such a philosophy troublesome lies in those instances where single-group or differential validity is considered. Under such circumstances, equally capable individuals from the group for which there are no valid predictors, or less valid than predictors for other groups, will have less chance of being selected. Under extreme conditions where a battery has little or no predictive utility for a group whose mean performance falls below the acceptable cut-off, no members of that group will be selected even if they are capable. Conversely, if the mean performance of a group on a battery that has no predictive utility for the criterion or criteria of concern falls above the cut-off, all members of the group will be accepted.

Qualified Individualism. This philosophy is very similar to unqualified individualism except in one major respect. Those who hold a philosophy of qualified individualism advocate the use of the best predictor or set of predictors except those that specifically identify an individual's group membership. When there is no systematic error (i.e., bias) in the predictors, then there would be no need to identify group membership as a predictor since it would add nothing to the prediction. However, if tests are biased and thereby predict differently across groups, the addition of group membership as a predictor may increase the test's utility, but would be objectionable to those who hold this philosophy. The major reason for the objection lies in the fact that such predictors are viewed as only being correlates to psychologically meaningful variables. Thus, they are considered a "stand-in" for the substantive psychological differences that exist among people. It is argued by those who hold a philosophy of qualified individualism, that a variable such as race is only related to the criterion or set of criteria in an obscure way. Its value as an explanation is remote at best and its use provides an easy vehicle for being lax in the quest for psychologically meaningful predictors.

Since group membership can not enter into the selection procedure in any way, then no adjustments to a biased test can be made. Additional predictors can be employed, but if this is done, it would have to be done for all since group membership can not be identified. This qualification would also rule out the use of different tests for different groups because, again, this would require treating groups differently soley as a consequence of a factor (i.e., race) that has no instrinsic psychological relevance.

The advantages and disadvantqes of this approach are similar to the philosophy of unqualified individualism with the following exceptions. First, qualified individualism has the additional disadvantage that since it cannot use "stand-in" predictors, there is more of a chance that assessment batteries will show differential validity across groups. On the positive side, the assessment practices evolving from a philosophy of qualified individualism can never be criticized for overtly making group membership a feature of the process.

Quotas. A quota philosophy is one in which maximizing predictive validity is seen as less important than adjusting cut-off scores to favor one or more groups. Such adjustments would allow a lower predicted criterion score for some groups and not for others. These adjustments to cut-off scores can be made for a variety of reasons. The various selection models designed to reflect this philosophy embrace a variety of values regarding fairness. All, however, disagree that selection based solely on predicting the same criteria cut-off score for all groups is the fairest of procedures.

Two types of selection models have been proposed under the quota philosophy. The first type argues that since the ultimate goal of decision making is to choose those who will succeed, the best way to set the cut-off score on the criterion is to base it on the potential success rate of different groups and not on what a validity study predicts performance will be on the criterion measure. Thus, this type adjusts for what is perceived to be unfairness when imperfect tests are employed. The second type quota selection model adjusts the level of predicted criterion performance for different groups based on the deemed social importance of having more of one group selected. It is regarded by its proponents as being fairer in that it regulates circumstances in a way that is believed to bring about "social good." In either of the two cases some applicants are selected that one would predict lower criterion performance than some applicants who are registered.

Selection Models

Several selection models have been proposed that reflect the various philosophies described above. Those that reflect the philosophies of unqualified individualism and quotas can be employed with either biased or unbiased tests. In the former, the different predictive utilities are corrected in the prediction equations for the various groups for which it has differential validity. Once accomplished, the same criterion performance level is used to determine the differing cut-off points on the predictor test(s) to use for each group. In the quota models, no adjustments are needed when biased tests are employed since a defined number of individuals from each group will be selected. Concern is geared toward the best ranking within groups. In the quota system differing cut-off points are chosen to both accomodate the bias in the test as well as provide the advantages to those groups as is deemed fair by the decision-makers. In most cases, it would be inappropriate to employ biased tests when one holds a philosophy of qualified individualism since any adjustments made to the scores would identify group membership. The only time it would be appropriate is if tests that were biased for one group were included with other tests that were biased for other groups in such a way that the biases balanced out. Thus, everyone would take all tests and no one would be identified by group membership.

In the remainder of this section, those selection models that have been most widely debated in the literature will be briefly described. It is not our purpose to provide detailed information about each model. For the reader wishing to employ one of the various models, we refer him/her to the references cited in this section. Thus, it is our purpose to provide an overview of the models and to classify them such that the reader may 1) become familiar with the various models that have been proposed and the philosophies governing their use, and 2) preview the more prominent models so the reader may decide on which one(s) he/she may wish to investigate further.

Equal Risk Regression Model. This model, named by Jensen (1980), is the simplest of models. It employs the same regression line in predicting criterion performance for all groups and the same criterial cut-off score is used for all groups to determine who gets selected and who doesn't. Since the same regression line is used, this model is only employed with unbiased tests. This model of selection is acceptable to those who hold unqualified and qualified individualism philosophies.

Equal Risk Model. As proposed by Einhorn and Bass (1971), this model first sets the minimum acceptable criteria performance and then chooses those individuals based on the maximum degree of risk of selection error decision makers are willing to tolerate. This risk factor is the same regardless of group membership. For each individual the best predictor or set of predictors is employed. Any test or set of tests may be used and different tests for different groups are acceptable. Employing the best predictor(s) for an individual, criterion performance is predicted and with the aid of a normal curve, the applicant's risk of failure is computed. Individuals, regardless of group membership, are selected if their risk is less than set as maximally acceptable. This model can be used for tests that are biased in slope, intercept or standard error of estimate, and is acceptable only to those who hold a philosophy of unqualified individualism. Since it identifies group membership in its selection of predictors, it is not an acceptable model to the qualified individualist. Those who hold a quota philosophy of fairness likewise find the model unacceptable in that the same criterion performance and consequent minimum acceptable risk are set the same for all individuals regardless of group membership.

Regression Model. This model, proposed by Cleary (1968), requires that for a test to be used, it must have the same slope and intercept for all groups. Once employed, the criteria cut-off predicted from the test is the same regardless of group membership. The distinction between this and the Equal Risk Regression Model is that there is no requirement that the test have equal SE_y across groups. As a consequence, the risk of failure can vary across groups if the SE_y is different for different groups. This is the case even though the prediction of criterion performance is the same across groups as the consequences of the equal slopes and intercepts requirement. Since identification of group membership is not necessary in selection, the model is appropriate for those who hold a philosophy of either unqualified or qualified individualism. The use of the same criteron cut-off, regardless of group membership, denies its acceptability for those who hold a quota philosophy.

Multiple Regression Model. Proposed by McNemar (1975, 1976), this model is most closely aligned with a philosophy of unqualified individualism. From the view point of unqualified individualism, it is the most statistically sophisticated and appealing of all models proposed to date (Jensen, 1980). The purpose of the model is to make use of the best possible set of predictors in the selection process. Group membership is used to statistically adjust for bias in predicting the criterion when systematic error is evidenced in the prediction. Consequently, it does not fulfill the requirements of qualified individualism.

Once the best predictions are made, the user of this model maximizes the average level of performance of the selectees by ranking them in order of predicted performance on the criterion and choosing those who either meet a minimum level or until the desired number of individuals has been selected. This is done without consideration of group membership and is consequently not an acceptable model for those who hold a quota philosophy.

Proportional Representation Model. Simply stated, this model holds as a requirement that the proportion of selectees defined by a criterion, such as race, be set equal to some preestablished proportion such as that represented in the United States population. Individuals are ranked within each group and the number of individuals chosen from each group is accomplished by selecting from the top down. This model, therefore, employs different test cutoff scores for different groups with consequent differences in predicted criterion performance across groups. As such, it can be classed as a quota model and is unacceptable to those who hold philosophies of unqualified or qualified individualism.

Culture-Modified Criterion Model. This model, proposed by Darlington (1971), explicitly identifies the decrease in minimally predicted criterion performance that decision-makers are willing to accept when selecting minority individuals. This is, in practice, accomplished by reducing the prediction of the criterion score of the nonminority group members to equate them with the predicted score of the minority group members. Such a practice, then, builds a desired "bias" into the test by changing the intercept by a predetermined constant. The degree of bias built into the formula must be determined by the tolerance of the decision-makers for accepting some applicants whose test performance would predict a greater probability of failure than others who are rejected. By using this model, the decision maker is forced to make explicit in the selection formula all considerations regardless of whether or not the considerations were to redress past injustices or to compensate for perceived biases in assessment practices yet to be identified. When employing biased tests, differences in cut-off scores would not only account for differences when adjusting for bias but also adjust for "other" considerations. The results of employing this model satisfies a quota philosophy since it accepts different levels of predicted performance across groups. For the same reason, it is unacceptable to those holding unqualified or qualified individualism in philosophies.

Constant Ratio Model. Proposed by Thorndike (1971), this model argues that when the mean difference between group scores on the predictor test(s) are greater than the mean difference between group scores on the criterion test unfairness occurs. Since the correlation between the test and the criterion is imperfect, there is the possibility that the above will occur. When it does, the cut-off point used for the predictor test may exclude from selection some of the low scoring group who would be expected to pass if previous group performance on the criterion were used to predict future performance. As a consequence, Thorndike suggests that the proportion of minority and nonminority group individuals selected should be the same as the proportion of the groups that would succeed on the criterion if given a chance. For example, if 30 percent of the minority group members and 40 percent of the majority group members succeed on the criterion, then the cut-off score on the predictor test should be set so 30 percent of the minority group members and 40 percent of the nonminority group members are selected. Since this model varies the acceptable criterion level across groups, it is a quota model. Yet, it differs from those quota models mentioned above in that the adjustments are made as a consequence of potential unfairness due to imperfect predictors, not based on ethical or moral values concerning the ultimate "good" of the selection process.

Conditional Probability Model. Similar to Thorndike's model, this model, described by Cole (1973) is a quota model based on a belief in fairness stemming from problems evolving out of the use of imperfect tests. Cole (1973) argues that there should be the same probability of selecting minority and nonminority group members as defined by each group's probability of achieving satisfactorily on the criterion. As the name implies, it differs from Thorndike's model in its use of conditional probabilities rather than constant ratios. However, its intent is the same.

Equal Probability Model. This model, proposed by Linn (1973) and named by Petersen and Novick (1976) is a quota model designed to equate chances of success across groups. Fairness is accomplished when all those who are selected have an equal chance of success, regardless of group membership. This requires that the predictor test cut-off scores be set so that the same proportion of minority group members will be selected who are predicted to succeed as nonminority group members. Under circumstances where there is a large discrepency between the means of the minority and nonminority groups on the predictor test with the minority mean below the nonminority mean and a high-cut off on the criterion, it would be necessary under this model to deny selection of some of the best nonminority applicants so that the proper proportions can be maintained.

Probability Weighted Model. This model was first described by Berieter (1975) and gives everyone some chance of being selected. However, the probability of their test being selected is defined by their probability of success as indicated on the predictor test. Under all of the other models discussed, there is a proportion of individuals whose performance would not allow them to be considered for selection. Berieter argues that because of the imperfect nature of tests even low scoring individuals have some chance of succeeding. Consequently, those individuals also should be considered, regardless of how small the chances are for selection. Making use of the cut-off score, the predicted score, and the SE_y, one can calculate the percent chance an individual has of succeeding. If one individual has one percent chance of succeeding and another 50 percent, then the latter individual should be given a 50 times greater chance of being selected than the former individual. The Probability Weighted Model is a quota model "since it selects from each group a proportion that equals the proportion that would exceed the criterion cut-off" (Jensen, 1980, p.407). In that sense it is similar to the Constant Ratio Model. However, while the Constant Ratio Model selects persons within groups to maximize the criterion performance of those selected, the random selection procedure of the Probability Weighted Model does not result in such maximization. Instead it allows for some who have a lower chance of succeeding on the criterion to be selected.

Expected Utilities Model. The applications of this model, first proposed for decision making in economics by von Neumann and Morgenstern (1944) and again by Wald (1950), has recently been explicated for use as a selection model by Gross and Su (1975), Petersen (1975), and Petersen and Novick (1976). This model is highly recommended by its proponents since it can be adopted by all decision makers, regardless of their fairness philosophy. The model forces the decision makers to decide explicitly what considerations, if any, they wish to include in the process. Then, weights are given to the desirability or utility of the various possible outcomes in such a way as to maximize the utility of the selections made. When we consider the fact that we are predicting performance on a criterion with less than perfect tests (i.e., tests without perfect reliability and validity) we can view the outcomes of decisions made with these tests as one of four types. These include situations where an individual is (1) selected and performs as predicted (true positive); (2) selected and does not perform as predicted (false positive); (3) not selected and would have performed as predicted (true negative); and (4) not selected and would not have performed as predicted (false negative). If each of these outcomes is given a weight, then the outcomes that are desired for each group tested can be decided beforehand and a formula constructed to meet those ends. Weights can be assigned to the desirability or utility of each of these outcomes and a selection formula constructed that would manipulate the probabilities of selecting each type. For example, if one wishes to use a quota to maintain proportional representation of minority and nonminority individuals, one can do so by varying the weights assigned to the outcomes across groups. If proportional representation is desired, then one may wish to give added weight to the outcome of selecting minority individuals for whom one may not predict success but may succeed given the chance. By manipulating this number, proportional representation can be assured.

As mentioned above, all philosophies of fairness can be satisfied through using this model. In addition, all models discussed in this section can be viewed as a derivative of this model since all require, either implicitly or explicitly, the assignment of utilities. In qualified and unqualified individualism the utilities across groups are the same. Each of the quota models varies utility across groups to effect fairness . For example, the Constant Ratio Model, when conceived from the above perspective, requires that the sum of the true positive and false positive expected utilities divided by the sum of the true positive and false positive expected utilities be the same across groups. The expected utilities are the assigned weights or utilities for each outcome multipled by the conditional probabilites of each outcome summed over all applicants. Likewise, the Conditional Probability Model, when stated in terms of the Expected Utilities Model, requires that the expected utilities of the true positive divided by the sum of the true positive and false negative expected utilities be the same across groups.

Fairness or Bias. When writing about the various models and their use in the selection process, some authors have referred to it as an issue of bias while others have referred to it as an issue of fairness. For our present purpose, we employ the use of both terms in differentiating between the two. In our review we have noted that bias can be equated with the concept of validity when validity is conceived in its broadest sense. We also implied that there are two classes of validity, construct validity and outcome validity, and consequently two forms of bias, construct bias and outcome bias. Outcome validity is that type validity which provides information on the utility of a test in predicting desired outcomes. Outcome bias, then, relates to tests that predict desired outcomes differently across groups. The choice of selection models briefly described above is not an issue of validity. The model chosen for use in any selection process is based on one's philosophy of fairness. There is no one philosophy that is more valid than another. However, whether or not the model when implemented results in the desired outcome (e.g., proportional representation) is an issue of outcome validity. If, when implemented, a systematic error results, then the use of the model is biased. From this perspective a fair model may be biased in the sense, and only in the sense, that its implementation does not yield the outcomes as predicted from the model.

Empirical Studies in Selection Bias. A few studies have been conducted that qualify under our present conceptualization of bias in selection since they focus on the validity of the test with respect to selection outcomes as opposed to the validity of the test in demonstrating it to be an effective measure of a construct. As mentioned previously, external criteria used in many predictive validity studies are chosen to demonstrate that the test, or a measure of a construct, is acting as the construct it is supposed to measure. While these studies provide the decision maker some information relevant to the situation in which they intend to use the test, the possibility exists that in certain situations the validity of the criteria falls short for specific decision making. The exception to this general rule lies in the employment testing literature. In this literature criterion-related validity researchers often choose criteria whose face validity is quite high for decision making. However, even in this literature concern has been raised that the general use of cognitive abilities tests to predict job performance may be requiring inferences that invalidate their carte blanche use across employment settings. Specifically, concern has been voiced that the predictive validity information on certain type cognitive abilities tests in predicting certain types of job performance does not warrant the generalized use of all cognitive abilities for tests for all job functioning. It may be that the validity of a test is situationally specific. Ghiselli (1966), after observing considerable variability in validity coefficients across studies, noted this concern. Schmidt, Hunter, and Urry (1976) examined this possibility and noted that tests that show validity as offered in one situation appeared to be invalid in up to 50 percent of the studies employing its use in predicting job performance in other situations. However, a recent series of studies have found this invalidity to be a statistical artifact mainly resulting from sampling error, differences across studies in test and criterion reliability, and differences in range restrictions (Callender & Osburn, 1980; Lilienthal & Pearlman, in press; Pearlman et al., 1980: Schmidt, Gast-Rosenberg & Hunter, 1980; Schmidt & Hunter, 1977;

Schmidt, Hunter & Coplan, 1981). In another study conducted by Schmidt, Hunter and Pearlman (1981), the use of tests that measure various cognitive abilities were found valid in predicting job performance across a family of five different clerical positions. Similar findings are reported by Hunter (1980) who showed that across 500 criterion-related validity studies employing a variety of criterion measures puported to be valid across a variety of jobs, the validities of a composite of verbal and quantitative ability measures in predicting class of jobs grouped according to their complexity of information - processing requirements, ranged from .23 to .56. Hunter (1980) concludes that there appears to be validity in using these type cognitive tests in predicting job performance even for the lowest skill jobs. As a result of these studies reporting on the generalizability of a variety of cognitive ability measures for a variety of job families, Schmidt and Hunter (1981) conclude that, with respect to employment testing "our evidence shows that the validity of the cognitive tests studied is neither specific to situations nor specific to jobs" (p. 1133). Another conclusion that can be drawn from these studies is that since they are unbiased in predicting job performance across groups in those studies, there should be no reason to question their unbiased nature when the tests are employed in a general way across situations and jobs.

In the application of tests for making educational decisions regarding special education diagnosis and placement, the validity issues are somewhat different. Tests validated for making such decisions are validated to demonstrate their utility in measuring a construct. External criteria used to validate IQ tests, for example, are chosen to demonstrate the validity of the test as a measure of intelligence. Similarly, when the same external criteria are used to determine if bias exists, the question addressed relates to whether or not the test is differently valid in the measurement of the construct across groups.

External criteria employed to validate IQ tests (usually a standardized measure of academic achievement) are related to a desired outcome of the selection phase of the decision making process (i.e., choosing those who will not perform without intervention). From this it is inferred that the predictive validity studies so offered provide validity for the use of the test in decision making. Similarly, with respect to bias, the assumption is made that if they are unbiased in measuring the construct, they are unbiased when they are used in decision making. However, whether or not an IQ test predicts if a child will or will not be able to perform with or without intervention equally well across situations for culturally different children is a question yet to be answered. Indeed there are those that would argue that not only has the question not been answered, but neither has the more basic question: "How well does an intelligence test predict that culturally different children will not perform differently if they are not selected for placement?"

It is pointed out by some that this question would have to be answered by using criteria more relevant to the decision making process than scores on standardized tests of academic achievement. Only a handful of studies are reported in the literature that address this concern. One such study, conducted by Goldman and Hartig (1976), was given an inordinate amount of weight in Judge Peckham's decision in the Larry P. case (see Chapter 8) for the very reason that the criteria to which it predicted was judged more closely related to that required for making EMR placement decisions than standardized tests of intelligence. This study produced quite different results than those reported in Chapter 4 under "External Construct Bias". Most evidence that examined IQ tests for differential validity in predicting academic achievement across races found no such evidence. In the Goldman and Hartig (1976) study, the authors employed a criterion measure of achievement grade point average (GPA) that included, among other school subjects, grades in music, health, art, and physical education. Correlations between the WISC Full-Scale and IQ and GPA were .25($p<.01$) for white children, .12 ($p<.05$) for Mexican-American children, and .14 ($p<.01$) for black children.

The correlation for whites is substantially lower than those reported in other studies, and the substantially lower correlations for minority children suggests that IQ tests may be invalid for use with other than nonminority populations in predicting school achievement (as differentiated from academic achievement). The Goldman and Hartig (1976) study, however, has several serious methodological flaws. First, GPA for blacks and Mexican-Americans showed considerable restriction of range. Second, groups were combined across schools and assumed to reflect a common standard used in grading. Consequently one must be concerned with the heterogeneity of the data collected on the criterion.

In another study of the relationship of the WISC-R factor scores to a 10-item teacher rating of academic performance, Reschly and Reschly (1979) obtained results more comparable to external construct bias studies employing standardized achievement tests than those found in the Goldman and Hartig (1976) study. In the Reschly and Reschly study, the correlation between the Verbal Comprehension factor of the WISC-R and teacher ratings were .30, .46, and .32 for whites, blacks, and Mexican-American students, respectively. The correlations between the Perceptual Organization factor of the WISC-R and teacher ratings were .22, .26, .27 for whites, blacks, and Mexican-Americans, respectively. The magnitude of the relationships were not as high as those relating IQ to standardized tests of academic achievement but they are similar in that they do not differ across groups. These relationships did not hold for Native American Papago students.

Using similar GPA criteria to that employed by Goldman and Hartig (1976) and teacher ratings of competence, sociability and social conformity, Mercer (reported in

Mercer, 1979) reports consistently higher correlations between WISC IQ scores and GPA for whites than for blacks and Mexican-Americans and from .26 to .28 for blacks. However, correlations between the GPA and teacher ratings and the Verbal Scale of the WISC were higher than the criterion measures and the Performance Scale of the WISC.

The results of the Mercer study as well as those of Goldman and Hartig are suggestive at best. No comparisons between the validity coefficients were reported in either study so it is impossible to determine if the reported differences are statistically significant. Additionally, the correlations between the WISC and GPA are of different magnitude (the Mercer correlations appearing somewhat higher) suggesting possible differences in the criterion measures used. When comparing the teacher rating studies of Mercer and Reschly and Reschly, the findings appear to be similar with no consistent differences appearing between groups. Surely, the paucity of research in this area leaves us wanting.

The question has been raised in the literature regarding the legitimacy of using school achievement as a criterion measure as opposed to a measure of academic achievement (Reynolds, 1982). Hopefully, the conceptualization of the various issues presented here, provides an alternative way of viewing this problem. When one is using an IQ test to measure intelligence, then certainly a measure of academic achievement is best employed since one would predict that the more intelligent a child is, the more that child will achieve academically. When viewed from this perspective, the use of a measure of school achievement that includes achievement in music and physical education, contaminates the purity of the criterion measure. However, intelligence tests are not used for diagnosis alone. They are also used for making placement decisions and sometimes for helping to design specific interventions. With respect to the former, the purist may argue that the placement is automatic once the classification is made. In practice that may not always be the case. Indeed, our personal observations suggest that often just the opposite is true, especially in cases where the diagnosis is unclear. That is to say, decision makers may first decide if the placement will benefit the child and then decide, according to their placement decision, whether or not to diagnose the child EMR. Legal requirements in some states mandating proportional representation also influences the diagnosis-placement decisions. Whether or not a district has met their quota of one group of children in EMR classes may also influence diagnosis.

Many other examples can be cited where the proverbial tail wags the dog. This dilemma is fed by the continuing requirement that diagnosis be a prerequisite to placement, a requirement with which schools sometimes find difficult to adopt especially given its unimportance in meeting their major purpose, helping the child learn better. The point is, that the decision to place a child is more complicated than just diagnosing him/her. When a decision maker is required to make a decision as to whether or not a child is in need of help, what information should they have on hand? Should it be information on predicting future performance on a standarized achievement test, GPA, teacher ratings or some other criteria? Should we not worry about making such decisions and only concern ourselves with making proper diagnosis? These questions can only be answered by those having to make the decisions after a critical analysis of the whole purpose of assessment activities.

Intervention Bias

The major purpose for employing tests in selection is to answer a question regarding an individual's future performance as predicted from the test. The use of tests for making intervention decisions, on the other hand, focuses on predicting an effective intervention from the test. While data gathered for selection decisions can tell us whether or not a child needs help, it is data gathered for intervention decisions that aid in identifying how to provide help. With respect to bias, a similar distinction can be made. Selection bias is bias that occurs when employing tests that result in systematic error in the identification of children across groups who need help while intervention bias involves systematic error in predicting successful interventions across groups. So, for example, a placement intervention (e.g., special education placement) that is effective for one group and not for another would be considered intervention bias. When we view it from within the conceputalization of validity in outcomes, intervention bias would be that which occurs in tests or other assessment strategies that are valid in predicting successful intervention for one group and less valid or invalid in predicting this desired outcome for another group.

As mentioned earlier in this chapter, data employed in decision-making can come from a variety of sources. Data can be generated from test-and nontest-based assessment strategies. These assessment procedures that generate both test- and nontest-based data have in common the fact that they are planned. We can distinguish the data generated from these planned procedures from data that are employed in decision making but not planned. Data derived from clinical impressions, the nature of the referral problem, and naturally occurring characteristics such as race, sex and socio-economic status are examples of what we are presently identifying as unplanned. While data drawn from clinical impressions can be planned in the sense that they are consciously derived from either test-or nontest procedures, they are unplanned in that they are inferred from assessment strategies designed for other purposes. The common feature of all unplanned data is that they are impressionistic.

Intervention Bias With Planned Data. One of the most intrusive interventions that commonly occur in schools is placement in self-contained special education classes. Subsequent to such an intervention a child is assessed by school personnel who employ various assessment strategies to determine if placement is warranted. It would follow, therefore, that if special education placement is to be the intervention, then the assessment data employed to determine placement should be able to predict that the child would benefit more from placement than from no placement. Such a prediction assumes that the intervention benefits some children. In the case of special education placement, this assumption has been questioned. A review of the empirical literature on homogeneous grouping for instructional purposes does not support this asssumption (Samuda, 1975), and the whole question of the value of special education as presently conceived as a form of intervention has been an ongoing topic of discussion (see Hobbs, 1975). When intelligence test data are employed, for example, to support placement in a class for the mentally handicapped, inferences are being drawn from that data for which there is no outcome validity evidence. Since intelligence tests do correlate with academic achievement there is some outcome validity evidence to infer that the child needs help, but to take it one step further and say that from the use of the test one can predict the child will be better off if he/she is placed, has no support in the empirical literature.

Given such a circumstance the issue of outcome bias with respect to placement becomes a moot point. In order to show that tests are biased with respect to making placement decisions, one needs to show that there is differential success of placement across groups. If there is no success for anyone, then the assessment data employed to make the placement are not biased towards any one group, just ineffectual for all groups.

In addition to placement decisions, other forms of intervention for children experiencing learning and adjustment problems in school are typically recommended based on both test-and nontest-based data. As discussed in Chapter 2, the assessment strategies employed in such circumstances, as well as the subsequent intervention recommended, are largely influenced by the assessors beliefs regarding the nature of the problem. Quay (1973) identifies three conceptual models that influence an assessor's views of the educationallly handicapped child. The first involves a belief that the exceptional child suffers from a dysfunction in either their cognitive, perceptual, or motor processing capabilities. This process dysfunction view further holds that the dysfunctional processes are unremediable. Such a view results in intervention recommendations that attempt to bypass or compensate for the "damaged" process or processes.

The second viewpoint, the experiential defect view, involves a belief that the problems in processing evidenced in the child are the consequence of defects in the child's experiences that have left him/her with the present dysfunction. Remedial recommendations drawn from such a viewpoint center around efforts to directly intervene where defects exist to remedy the effects of deficient experience.

The third viewpoint holds that the problems encountered by the child are not a function of any defects or dysfunction within the child but rather by the lack of proper exposure or instruction. This third viewpoint, the experience deficit view, leads to assessment and consequent interventions that directly address the skill deficit evidenced in the child.

Consistent with these viewpoints have been a variety of assessment-intervention models proposed in the literature. Those who hold the first two viewpoints in which the problem is believed to be a problem in processing, have proposed a variety of diagnostic-prescriptive models to help children. The assessment techniques employed in these models, such as the ITPA, are designed to measure defective or dysfunctional processes. Those who hold the experience deficit viewpoint have proposed what Ysseldyke and Salvia (1974) refer to as the task-analytic or skills training approach. These approaches usually employ assessment techniques such as direct observation or criterion-referenced tests to measure specific deficit skills.

Ysseldyke and Mirkin (1982) identify a variety of diagnostic-prescriptive models that have been proposed to deal with a myriad of inferred processing problems. These include models designed to addres vision problems (Bernetta, 1962; Coleman, 1968; Coleman & Dawson, 1969; Debard, Houghton & Thomas, 1972; Ewalt, 1962; Forrest, 1968; Getman, 1962, 1966a, 1966b, 1972; Getz, 1973; Gould, 1962; Greenspan, 1973; Halliwell & Solan, 1972; Kane, 1972; Kirshner, 1967; Mullins, 1969; Rosner, 1963; Slota, 1971; Swanson, 1972; Swartout, 1972; and Tramonte, 1963), psycholinguistic problems (Kirk & Kirk, 1971; Kirk, McCarthy, & Kirk, 1968; Minskoff, Wiseman & Minskoff, 1972), problems in perceptual motor dysfunctioning (Arena, 1969; Barsch, 1965, 1967, 1968; Behrmann, 1970; Dunsing & Kephart, 1965; Early & Sharpe, 1970; Frostig, 1967, 1972; Frostig & Horne, 1964; Forstig, Lefever & Whittlesey, 1961; Johnson & Myklebust, 1967; Kephart, 1960, 1964, 1971, Magdol, 1971; Marten & Harmon, 1962; Roach & Kephart, 1966; Smith, 1968; Sutphin, 1964; Van Wetsen, 1967), sensory integration problems (Ayres, 1972), modality problems (deHirsch, Jansky & Langford, 1966; Lerner, 1971; Johnson & Myklebust, 1967; Wepman, 1967), and problems in rhythm and body balance (Rice, 1962).

Programs identified by Ysseldyke and Mirkin (1982) designed to represent a task analytic or skills-training approach include directive teaching (Stephens, 1976), direct instruction (Carmine & Silbert, 1979), DISTAR (Becker & Engelmann, 1978), data-based instruction (Deno, 1972; Fox, Egner, Paolucci, Perlman & McKenzie, 1973), data-based program modification (Deno & Merkin, 1977), exceptional teaching (White, Haring, 1976), individual instruction (Peter, 1972), precision teaching (Lindsley, 1964, 1971), and responsive teaching (Hall & Copeland, 1971). In addition to these general intervention models, the behavior therapy literature is robust with additional skills-training approaches. While the above models are general models described as useful for the assessment and intervention of a variety of problem behaviors, other behavioral approaches, although commonly employing similar principles of learning (e.g., reinforcement), are problem specific. While diagnostic-prescriptive models focus on efforts to remediate processes, the task-analytic or skills training models focus on their adherence to "sequential, systematic, intensive, individualized or small group instruction on skills that are directly related to the academic and social requirements of the school program" (Ysseldyke & Mirkin, 1982, p.398).

Empirical literature on the outcome validity of the assessment approaches employed in the diagnostic prescriptive and skills training models is revealing. In reviewing the literature on diagnostic-prescriptive models, Ysseldyke (1973) describes three common research methodologies that have been employed: (1) descriptive, (2) gain-score, and (3) aptitude-treatment interaction (ATI). The first, descriptive, attempts to establish a relation between the ability or process and academic achievement. Such information lends towards the validity of the construct and selection validity of these tests. With respect to this descriptive research, Ysseldyke and Mirkin (1982) conclude that "in spite of numerous textbook claims for the relationship between performance on measures of specific abilities and on measures of academic achievement, extensive reviews of the research indicate little empirical evidence for such claims" (p. 400).

Gain-score research that attempts to show gains in ability and/or academic performance after the institution of ability training provides little or no empirical support for a variety of programs examined. Most heavily researched are psycholinguistic and perceptual-motor training programs. Unfortunately, research in this area is characterized by serious methodlogical flaws. Failure to consider the Hawthorne effect, regression effects, linearity across different levels, and lack of reliability in the measures employed in the assessment of both ability and achievement make interpreting this literature extremely difficult (Ysseldyke, 1973). Evidence from the methodologically sound gain-score studies provides little support for the validity of these interventions.

ATI research employs a sound methdology for examining the effects of intervention program with efforts to identify the differential effect of instructional treatments with children who differ on certain abilities. The goal of the research in this area is to show that individual differences (e.g., intelligence) are important to consider when designing instructional programs. So, for example, an interaction between the various levels of an attribute across individuals and the treatments employed would lend evidence for prescribing different treatments to those who differ in the attribute. Research evidence in search of ATIs have met with little success. In a review of 90 ATI studies, Bracht (1970) found 85 of them to produce no predicted interactions (i.e., disordinal). Mann, Proger, and Cross (1973) have concluded that the main problem with ATI research appears to be in the difficulty with adequately measuring the aptitutde under investigation.

It can be concluded from the abundance of evidence available to date that there is little empirical support for predicting effective interventions from the process tests commonly employed in special education decision making. With respect to interventon bias, then, we again find ourselves in the position of suggesting that there is no evidence of intervention bias with diagnostic-prescriptive approaches for the simple reason that there is no support for their validity with any group.

The literature on skills training approaches have been more successful in demonstrating the effectiveness of interventions. Consequently there is a literature that lends outcome validity evidence for the use of the assessment strategies employed in these approaches. Consistent with the experiential deficit view of educational exceptionality, these assessment approaches are direct in that they focus on the measurement of behaviors that are directly related to the presenting problem. This is in contrast to the behaviors measured in the diganostic-prescriptive approaches that are inferred to be indirectly related to the presenting problem. So, for example, if the presenting problem is poor reading achievement, the skills training approaches focus on behaviors that are functionally related to reading while diagnostic-prescriptive approaches focuses on the measurement of behaviors that are signs of a process which in turn is inferred to be related to reading achievement.

The most successful of the skills training approaches have been those that employ continuous measurement of the targeted criterion behavior. Given the tentative nature of our understanding of the assessment-intervention process it is rather presumptuous to assume that a single measure taken before the establishment of an intervention can provide the information necessary to plan and implement an effective intervention (Deno, Mirkin & Shinn, 1978). The continuous collection of data allows for continuous refinement in the intervention program and consequently more effective learning (Van Elten & Van Elten, 1976).

Programs employing direct and continuous measurement of performance and the use of these data to make corrections in subsequent programs have considerable empirical support (Ysseldyke & Mirkin, 1982). The major intervention component in these programs in continuous measurement itself. In addition, these interventions commonly employ reinforcement and feedback. Two such programs, namely, precision teaching and data-based instruction have been particularly successful in addressing math and reading behaviors (Bohannon, 1975; Bradfield, Brown, Kaplan, Rickert & Stannard, 1973; Deno, Chiang, Tendal & Blackburn, 1979; Haring & Krug, 1975; Haring, Maddux, & Krug, 1972; Mirkin, 1978; Mirkin, Deno, Tindal & Kuehale, 1980). Findings from these research efforts point to the importance of utilizing the continuously collected data to effect changes in the intervention. This has been most effectively accomplished through the establishment of decision rules based on the success of daily goal attainments (Liberty, 1975). The use of students to grade and graph their own progress has also been shown to be an effective method for utilizing continuously collected data (Fruness, 1973).

The skills training models while demonstrating intervention validity for the assessment procedures employed, have not addressed the issue of intervention bias. There is no evidence reported in this literature bearing on the differential impact of the interventions across groups. Consequently, the potential intervention bias of the assessment procedure has yet to be determined.

Bias with Unplanned Data. The decision-making process is a complex one that draws information from a variety of sources in reaching decisions. Some of the data used in the process have validity for predicting the outcomes of interest while others do not. Likewise, some of the data employed in decision-making are biased in that they predict outcomes differentially across groups. In the last section of this chapter we saw how planned data from only a limited number of procedures have been demonstrated to have outcome validity with respect to intervention planning and of these procedures, more has been empirically studied to determine intervention bias. In this section we turn attention to the potential outcome bias in the development of interventions with unplanned data. As mentioned earlier, unplanned data are data based on the impressions of the decision-maker(s) and is drawn from direct or vicarious experience with a child or children believed similar to the child under study. Clinical impressions, the influence of naturally occurring characteristics, such as race, sex, socio-economic status and attractiveness, and the impact of the referral problem on decision making can all be classified as unplanned data. These data can either directly or indirectly influence decision making and its inclusion in decision making can only be justified on the grounds that its use increases the validity of the decisions made. With respect to bias, its use would have to preclude differentially effective interventions across groups.

The literature on clinical impressions in this area is negligible. The few studies that exist provide no support for the use of clinical impressions in either diagnosis on treatment (kazdin, 1978). Several studies reporting the influence of naturally occurring pupil characteristics have recently been reported in the literature. It is the assumption of this literature that employing factors such as race, socio-economic status and physical attractiveness is inappropriate and a biasing factor in the decision-making process. However, whether or not the use of these factors results in intervention bias is an empirical question. There is a question of fairness, however, that is posed by the use of these factors. As discussed in Chapter 4, those who hold a philosophy of unqualified individualism consider their use fair as long as they increase validity. Those who hold a philosophy of qualified individualism find their use unfair, since, if they have any predictive utility, its only because they are correlated with psychologically meaningful variables. In other words, they have no intrinsic meaning.

Those studies that have specifically examined the influence of naturally occurring characteristics have attempted to identify the unconscious impact of these factors on special education decision making. Typically, these studies have manipulated race (Frame, 1979; Matuszek & Oakland, 1979; Tomlinson, Acker, Canter, & Lindborg, 1977), SES (Frame, 1979; Matuszek & Oakland, 1979; Ysseldyke & Algozzine, 1979), and physical attractiveness (Ross & Salvia, 1975; Salvia & Podol, 1975; Ysseldyke & Algozzine).

Research on the influence of race on decision making has not shown race to be a significant variable in influencing school psychologists' diagnoses (Frame, 1979). With respect to placement decisions, Frame (1979) found an interaction between race and SES but in an unexpected direction. In this study lower-class black children were less likely to be recommended for placement than upper-class blacks or lower and upper-class whites. In the Matuszek and Oakland (1979) study, race did not influence school psychologists' placement decision but SEX did. Consistent with Frame's study, lower SES children were less likely to be recommended for placement an upper SES children. In the Matuszek and Oakland study, teachers were not influenced by their placement decisions by either face of SES. In a study of the placement recommendations of black, Native Americans, Indians and Orientals, Tomlinson et al. (1977) similarly found that special educaton placement was more likely for whites than the minorities in their study. In addition, they found that minorities were more likely recommended for resource room placement. Ysseldyke and Algozzine (1979) also found that the participants in their computer-simulated decision making study (i.e., school psychologists, special and regular education teachers, administration counselors, nurses and social workers) reported that SES influenced their diagnostic decision making more when students were from high SES families rather than low SES families. However, SES had no measured impact on their actual diagnostic decisions. Reynolds (1982) suggests that the trend not to place lower SES black children in EMR special education classes may be consequence of psychologists' tendency to rate the "true intellegence" of these children higher than their performance indicates.

In a study examining the influence of physical attractiveness on diagnoses made by classroom teachers, Ross and Salvia (1975) found that less attractive children were more likely to be diagnosed mentally retarded than more attractive children. Similarly, Salvia and Podol (1975) found that speech therapists rated identical speech samples lower when they believed the sample was from a child with a repaired but visible cleft palate than when they believed it was from a normal child.

Basically, the literature on naturally occurring characteristics indicate that race, SES and physical attractiveness appear to be important variables for further study in special education decision-making. The fact that they are used raises three questions. First, are they valid? Second, are they biased in the sense that their use results in differentially effective interventions? Third, are they fair? The last question, of course, is not an empirical one.

The last factor receiving some attention is the empirical literature that can be identified as unplanned, in the influence of the type of referral on special education decision-making. Ysseldyke and Algozzine (1979) report that the diagnostic decisions regarding emotional disturbance were influenced by the referral in their study. The fact that a child was referred because of a behavior problem had an impact on diagnostic decision-making independent of the planned data that was provided the decision maker. No evidence is available regarding the validity of using the reason for referral as data for decision making on the potential bias of their use.

ARIZONA CENTER FOR EDUCATIONAL EVALUATION AND MEASUREMENT

Nonbiased Assessment in Psychology and Education

Vol. II

November, 1982

THE UNIVERSITY OF ARIZONA
COLLEGE OF EDUCATION
TUCSON, ARIZONA 85721

This project has been funded at least in part with Federal funds from the Department of Education, Office of Special Education, under Grant Number G008100160, Grant Authority CFDA: 84.023H. The contents of this publication do not necessarily reflect the views or policies of the Department of Education, Office of Special Education, nor does mention of trade names, commerical products, or organizations imply endorsement by the U.S. Government.

Nonbiased Assessment in Psychology and Education

Vol. II

November, 1982

[Final Report]

Thomas R. Kratochwill
The University of Arizona

Anthony A. Cancelli
Fordham University

Chapter 7

Proposed Alternatives to Traditional Assessment

As noted in previous chapters, many criticisms have been leveled at traditional testing and assessment practices. As the limitations of traditional testing practices became more salient and widespread, alternatives to these procedures emerged. In some cases the alternatives have a long history in their own right and have been examined for their utility as nonbiased measures only recently. In other cases, completely new assessment strategies have emerged, evolving out of completely new theoretical and conceptual models. Also, there has been a whole set of attempts to revise or modify the more traditional procedures to meet various definitions of nonbiased testing.

In this chapter some assessment techniques and procedures that provide alternatives to traditional measures are reviewed. This chapter provides a review of culture-reduced testing, renorming, adaptive behavior measures, Piagetian strategies, learning potential assessment, diagnostic clinical teaching, child development observation, neuropsychological assessment, and behavioral assessment. Relatively more attention is devoted to behavioral assessment because these strategies have rarely been presented as alternatives to traditional measures within the context of test or assessment bias. Included in the section on behavioral assessment is a discussion of the conceptually compatible area of criterion-referenced testing.

Within the psychoeducational assessment procedures outlined in subsequent sections of this chapter some unique features of child behavior assessment should be considered. These include the referral source, social control, concomitant social influences, and cognitive assessment (Evans & Nelson, 1977, pp. 603-605).

The first issue is that the learning and behavior problems of the child are typically identified and presented by a socialization agent (e.g., parent, teacher, physician). Thus, the referral may bear little or no relationship to the child's perception of the problem, although this may not always be the case.

Second, the young child is typically under much stronger and more obvious sources of social control than an adult or adolescent client, thereby requiring that the child's social environment be adequately and thoroughly assessed. Even when an exact organic etiology can be documented (e.g., definite and specific brain damage), a comprehensive environmental assessment should be completed.

A third consideration is that children with serious learning and behavior problems are very frequently already involved in some attempt to alter their behavior. Such alterations may occur prior to, during, or after some therapeutic intervention. For example, the psychologist conducting a behavioral assessment of a child experiencing reading difficulties may find that the child is receiving special attention from the classroom teacher, tutoring after school, or possibile involvement in one of many packaged remedial programs in the home. Such issues should prompt assessment of these programs and their contributions (positive or negative) to other interventions.

The fourth characteristic of child assessment is that it is frequently linked with cognitive or physical assessment. This position assumes that some learning problems are inexorably linked to physical or cognitive (intellectual) deficits. Thus, in addition to assessing a child's learning problems, assessment should usefully focus on vision, hearing, speech and other relevant physical problems.

Finally, developmental variables are intricately involved in assessing child behavior. Evans and Nelson (1977) indicate that "the younger the child the more must deviation from developmental norm be the measure of abnormality" (p. 605). Thus, the conventional notion of "spontaneous" recovery or improvement must be taken into account. For example, Schechtman (1970) has observed that some problems (e.g., poor concentration) show more pronounced decrease with age than others (e.g., poor schoolwork). Children experiencing fears and related problems may not need extensive intervention for successful elimination of the problem (Morris & Kratochwill, 1983). The professional must then consider that the social significance of certain behavior and/or learning problems will alter with age.

Culture-Reduced Testing

The notion of culture as it relates to bias in testing has received much attention due to its inherent implications in psychological assessment. Since psychological and educational measures are most often samples of static behaviors (i.e., behaviors sampled at one point in time) certain assumptions regarding the environmental or cultural impact on the acquisition of those behaviors are inherent in the design of any test. The conclusion that any measure of behavior must be understood as an interaction between one's genetic predisposition and the culture of the individual whose behavior one is attempting to measure has been recognized throughout the history of testing. In an attempt to minimize the impact of cultural bias, for example, Binet, in the development of the 1911 version of his intelligence test, scrutinized items to eliminate those he felt were overly dependent on information specific to the culture of the school or to that culture transmitted more readily in educated homes (Jensen, 1980). Despite the almost universal acceptance that any sample of test behavior is a product of an interaction between ones inherent potential and the culture from which he/she comes, much debate is still underway regarding the degree of the impact of the environment on behavior, the differences of the impact due to variation in culture, and the extent to which the content of tests reflects cultural differences that may result in bias.

A variety of terms have been offered in recent years to discuss the last of these topics. One of the first terms used in the literature to connote a test that was free of cultural influence was "culture free". While receiving a flurry of attention in the fifties this concept has since been discredited on the grounds that any test of mental ability must depend on some experiences acquired in some culture. Simply stated, a test can't be "free" of culture. In addition, attempts to design "culture free" tests did not meet with those results expected by their designers; that is, no cultural group mean differences in performance.

The concept of "culture free" soon gave way to the concepts of "culture fair" and "culture reduced". These conceptualizations were designed to connote tests or test items that "reduced" dependency on a particular culture for correct responding and, therefore, are "fair" to all individuals regardless of cultural background.

The Cattell Culture Fair Test of g and the Raven's Progressive Matrices are two such tests that have attempted to reduce the degree to which the culture in which one lives influences performance. Other tests such as the Leiter International Performance Scale, the French Pictoral Test of Intelligence and the Nonverbal Test of Cognitive Skills also fall into this classification. One of the basic aims of these type tests is to reduce verbal content. To varying degrees the amount of verbal instructions for the items and the amount of verbal responding required to correctly answer items are minimized. In addition, the content of the items is usually designed to reflect novel situations requiring the application of complex cognitive skills as opposed to situations heavily dependent on cultural experience. Such tests have been touted by some to be less biased against cultural minorities and a more acceptable measure of intelligence than traditional IQ tests such as the WISC-R and Stanford-Binet. Research evidence, however, indicates that mean differences across cultural groups is approximately the same in culture-fair tests as they are in conventional IQ tests (Arvey, 1972). This provides additional evidence for those who argue that the differences between groups on mental abilities tests are real differences in intellectual ability (Jensen, 1980).

Tests or test items that are not culturally "reduced" are often termed "culturally bound" or "culturally loaded". The degree of cultural influence on a test or test items can be viewed on a "hypothetical continuum" ranging from "culture free" to "culture bound" (Jensen, 1980). While this continuum provides for a simple conceptual understanding, distinctions along the continuum have in recent years become less important than the now more popular distinction between "cultural bias" and "cultural loading". Those who use the terms "cultural bias" and "cultural loading" when speaking of the influence of culture on tests and/or test items tend to devalue the importance of the hypothetical continuum, suggesting instead that it is a given that all items are culturally loaded. The degree to which an item or test is culturally loaded can not be an indicator of bias, since we have no criteria upon which to judge its place on the continuum (Jensen, 1980). The judgment of whether or not an item is "culturally biased" or simply "culturally loaded" is an empirical question and should be treated as such. Clarizio (1979), for example, argues that there has been much confusion in the intelligence testing literature due to a failure to distinguish between cultural bias and cultural loading.

Renorming

An alternative to designing tests with culture-reduced content in an effort to equate performances across groups is to use conventional tests such as the WISC-R and interpret the results so as to equate performance across groups after the fact. The content of the test remains the same for all, and the administration doesn't change. What changes is what one does with the data after it's collected. By far, the most popular attempt to employ this approach is the System of Multicultural Pluralistic Assessment (SOMPA; Mercer, 1979).

The SOMPA is a comprehensive system for assessing intellectual functioning that is based on a cultural and structural pluralistic model of society. Basically, Mercer (1979) argues that traditional testing practices are based on the Anglo conformity model of society. Tests based on this model assumes "that all children in American society are either being reared in families that have been both culturally and structurally integrated into the Anglo core culture, or are members of families of ethnic groups that are culturally integrated even though they may still be structurally separate and identifiable" (Mercer, 1979, p. 19). Mercer believes that such a view results in assumptions about test performance that are unfair. If all children are not integrated into the Anglo core culture, then tests based on assumptions of common cultural experience are going to be biased against those who are not. Accepting the view that we live in a pluralistic society, requires, according to Mercer, interpretive schemes that are capable of comparing performances only among children from similar backgrounds. The SOMPA, in part, is purported to do just that.

The SOMPA employs the use of ten measures within three assessment models: the medical model, the social systems model, and the pluralistic model. The medical model is employed to determine if the child is biologically normal. The six measures include Physical Dexterity Tasks, Weight by Height, Visual Acuity, Auditory Acuity, Health History Inventories and the Bender Visual Motor Gestalt Test. Two measures are employed to assess the child from a social systems model perspective: the WISC-R and the ABIC. The purpose for their use is to gauge how well the child is performing relative to his/her various social roles. Based on a model of social deviance, the measures attempt to identify if the child is abnormal in the sense that he/she deviates from the expectations of others in the group. Within this context WISC-R scores are perceived as achievement scores and labeled School Functioning Level (SFL). The purpose of the pluralistic model is to assess the child's learning potential. It employs the WISC-R for this purpose but adjusts scores according to the social and cultural group from which the child comes. In order to make the adjustments, the Sociocultural Scales are employed. These scales ask questions in four areas: Family Size, Family Structure, Socioeconomic Status, and Urban Acculturation. Three equations, one each for Verbal, Performance, and Full Scale IQs, are employed to estimate the learning potentials separately for Black, Hispanic, and Anglo children. A child is first classified as Anglo, Black, or Hispanic. Next, his/her scores on the Sociocultural Scales are computed and weighted in accordance with the group he/she belongs. Finally an Estimated Learning Potential (ELP) is derived. Similar to the WISC-R IQ, the Estimated Learning Potential scores have a mean of 100 and a standard deviation of 15.

The standardization of the SOMPA is based on a sample of 2085 California children representing equal numbers of Black, Hispanics and Anglo children. It is appropriate for use with children aged 5 to 11, inclusive. The various measures are taken in interview with the child's parent or guardian and through direct testing of the child.

While the SOMPA has been criticized for a variety of reasons, the most pointed criticisms involve its conception of the ELP. While the pluralistic model on which the measure is based is by no means controversial, the use of this model to derive estimates of learning potential is. Most would agree that all things being equal, the learning potential of children from various cultural groups would be equal. The major criticism arises from how and why the ELP is derived. Goodman (1979a), for example, points out that gathering a few facts about the SES, acculturation and the family of the child, is inadequate information to gain an accurate picture of a child's learning potential. Clarizio (1979) argues that the potential impact of using such a system will do more harm than good. By declassifying children who are now eligible for special education placement, many needed services will be lost. Goodman (1979b) also points out that the ELP carries with it the assumption that children whose scores are not raised by the adjustment to WISC-R scores are biologically impaired. She argues that this conclusion is not only dangerous but a throwback to the days when pure measures of biological potential were believed possible.

As these and other authors have pointed out, the true value of the ELP construct rests in its validity. In this regard, the ELP has little to support its use. The relationship of the ELP to academic achievement is lower than IQ to academic achievement (Oakland, 1979). Consequently, the use of the ELP in place of the IQ reduces ones ability to predict academic achievement. In defense of the ELP, Mercer has argued, as we have in this report, that predicting to standardized achievement tests may have its problems since both may be measuring the same thing, achievement, not potential. Consequently, the fact that ELP scores do not predict academic achievement does not automatically invalidate it. However, arguing that IQ is not a measure of potential and performance on academic achievement tests is not a measure of learning says nothing about the validity of the ELP. If the ELP has predictive validity then it must be demonstrated that it is related to some measure of learning. If not standardized achievement tests, then some other measure of learning. Similarly, if it is used for the purpose in which it was intended (i.e., educational placement), then the probability of those placed receiving effective intervention would have to be demonstrated, the predictions would have to be better than those derived from use of IQ tests, and it would have to be shown that its use would result in equally effective treatment across groups.

Adaptive Behavior Assessment

The measurement of adaptive behavior as a unique component in the assessment of mental retardation is a relatively recent occurrence. As Reschly (1982) points out, early definitions of adaptive behavior as exemplified in the American Association of Mental Deficiency Manual (Heber, 1961), placed heavy emphasis on learning ability as the criteria for evaluating adaptive behavior in school-aged children. Consistent with this conceptualization, adaptive behavior had traditionally been measured through the use of standardized achievement tests for this age group. Consequently, a child who performed poorly on a measure of intellectual functioning [e.g., in the borderline or mentally deficient[5] range according to Wechsler's classification system (Wechsler, 1974)] and, in addition, was performing poorly in school as defined by a measure of academic achievement, qualified as mildly or educably mentally handicapped. Recent definitions, however, provide a view of adaptive behavior that is multifaceted and require more than the measurement of academic achievement to adequately assess its various components. The 1977 edition of the AAMD manual (Grossman, 1977), for example, suggests that essential coping skills such as those involving the concepts of time and money, self-directed behaviors, social responsiveness, and interactive skills be included in measures of adaptive behavior. This change requires that more attention be paid than previously to the measurement of adaptive behavior. Given the judicial and legislative mandates for its use (see Chapter 8), those involved in the assessment of mental retardation have been searching for psychometrically sound methods of measuring adaptive behavior.

Despite this need, the lack of a clear specification of what constitutes adaptive behavior has resulted in the publication of a variety of different tests and procedures that all purport to measure adaptive behavior. Reschly (1982) identifies four features among the various definitions of adaptive behavior that are noteworthy for discussion; (1) developmental, (2) cultural context, (3) situational or generalized, and (4) domains. The developmental feature refers to the stated or implied understanding in all definitions that the criteria for assessing adaptive behavior changes with age. This is exemplified in the AAMD criteria (Grossman, 1977) that identifies sensory motor skills development as criteria during infancy and early childhood and not during childhood, adolescence or adult life. With regard to the second feature, cultural context, Reschly (1982) reports that most definitions of adaptive behavior acknowledge the importance of cultural influences on the development of adaptive behavior and recognize the need to interpret adaptive behavior within the context of the individual's cultural background. The third feature of adaptive behavior definitions refers to the dynamic nature of the construct. As mentioned above, current definitions of adaptive behavior are multifaceted, regardless to which age group one is referring. The question then arises regarding the relationship among the various facets or domains and the influence of different social systems and roles on each. Most definitions, according to Reschly, imply that the various domains are situation specific in that they are functionally independent sets of skills and that cultural background may influence differently the acquisition of each set of skills. The fourth feature addresses the issue of which domains are covered by the various definitions. Reschly reports that nearly all include the notions of self-maintenance or independent functioning, and language/communication skills, most include a reference to interpersonal competencies, and some include intrapersonal competencies (e.g., emotional maturity), and social responsibility (e.g., meeting social expectations such as constructive participation in the family and community). The various conceptions vary markedly on the weight ascribed to each of these domains. Likewise, the measures of adaptive behavior have emphasized different aspects of the concept.

Other characteristics that differentiate among the various measures of adaptive behavior are the purpose for which they are intended, the population for whom they are useful, the person or persons who provide the data, administration time, and their psychometric adequacy. Such information for most of the available adaptive behavior instruments has been presented by Oakland and Goldwater (1979) from which Table 7.1 is reproduced.

Coulter and Morrow (1978) distinguish two general purposes for which adaptive behavior measures are designed. The first is for diagnostic/ classification purposes (e.g., the Adaptive Behavior Scale for Children) and the second is for instructional planning (e.g., the AAMD Adaptive Behavior Scale). The former has gained importance as a consequence of the perceived bias resulting from the use of IQ and achievement tests alone in the diagnosis of mental retardation.

Adaptive behavior measures also differ on the intended population. The age range and severity of the retardation for which they are useful differ among the available measures (see Table 7.1). As a general rule, those that are designed for remedial instruction are for use with the low functioning mentally handicapped and cover wide age ranges. Those designed for classification/placement, are useful for mild mental retardation and are specific to certain age ranges.

Table 7.1 - Measures of Adaptive Behavior

	BEHAVIORS ASSESSED							ENVIRONMENT					Population type (and size)		Purpose			Examiner			Respondent			Reliability and validity data available	Scores			
	Phys. Dev./Sensory Motor/Locomotion	Self-direction	Language and Communication	Vocational and Occupation	Economic	Social	Self-Help/ Independent Functioning/Self-Maintenance	With Peers	In School	In the family	In the Community	Age Range	clinical	school	screening	placement	programming	teacher	diagnostician	para-professional	teacher	parent/family	child	Yes/No	G.E./Age	%ile	Scaled score	Administration time
AAMD Clinical (Nihira, et al., 1974)	X	X	X	X	X	X	X					3-adult	X			X	X		X	X		X		Yes		X		45'-1 hour
AAMD School (Lambert et al., 1974	X	X	X	X	X	X	X					7-13		X		X	X	X	X	X	X	X		Yes		X		45'-1 hour
Cain Levine (Cain, et al, 1963)		X	X			X	X					5-13				X		X	X	X		X		Yes		X		20'
California Preschool (Levine, et al., 1969)			X			X						2-5			X			X	X	X	X	X		Yes		X		20'
Camelot (Foster, 1974)	X	X	X	X	X	X	X					2-adult				X	X	X	X	Xa	X	X		Yes				1 hour
Adaptive Behavior Inventory for Children (Mercer & Lewis, 1978)				X	X	X	X	X	X	X	X	5-11		X		X			X	Xa		X		Yes		X	X	1 hour
Preschool Attainment Record (Doll, 1966)	X	X	X			X	X					birth-7				X		X	X	X	X	X		No	X		X	20'
Vineland (Doll, 1965)	X	X	X	X	X	X	X					birth-25	X			X			X		X	X	a	Yes				20'

a With extensive training in interviewing.

Reprinted from Oakland and Goldwater (1979, p. 147). Reprinted with permission.

BEST COPY AVAILABLE

ED 246 637 NONBIASED ASSESSMENT IN PSYCHOLOGY AND EDUCATION. VOLUMES I AND II.
[FINAL REPORT].
4 OF 7 KRATOCHWILL, THOMAS R.; CANCELLI, ANTHONY A. ARIZONA UNIV., TUCSON.
COLL. OF EDUCATION. NOV 82 603P.

Almost all adaptive behavior measures employ either a parent, guardian, or teacher as respondents (see Table 7.[1]). Only one measure, Children's Adaptive Behavior Scale (Richmond and Kicklighter, 1981), is designed as a paper and pencil test for collecting data from children and one, the Vineland (Doll, 1965), may be used for interviewing the client if the assessor is extensively trained. As reported in Table 7.1, most of the measures of adaptive behavior range in the time they take to administer from 20 minutes to one hour.

Given the recency of the restructured concept of adaptive behavior and measures conceived to reflect this concept, it is easy to understand the limited availability of validity evidence. While some of the measures have adequate construct validity, there is very little research reporting on the outcome validity of the measures. There has been some research reporting on the technical test bias of these instruments and the little research that has been conducted on outcome validity has focused on outcome bias. Reviewed below are two popular measures of adaptive behavior, measures we judge to be the most psychometrically sound for use in special education decision-making.

AAMD Adaptive Behavior Scale - School Edition (ABS-SE)

The ABS-SE (Lambert, 1981) and its predecessor, the Adaptive Behavior Scale - Public School Version (ABS-PSV; Lambert, Windmiller and Cole, 1975), were designed from the original AAMD adaptive behavior scale (i.e. Adaptive Behavior Scale - Clinical Version; Nihira, Foster, Shelhaas and Leland, 1969) developed at Parsons State Hospital and Training Center in Kansas with the support of NIMH. This original scale and its 1974 revision were designed for instructional planning for the severely retarded. The ABS-PSV was an offshoot of the ABS-CV and developed to help in both the instructional planning and classification/placement of children in special education classes. The ABS-PSV retained most of the items from the ABS-CV, eliminating those that could not easily be answered by teachers who are typically used as third party respondents. The revised ABS-SE was developed in "response to the need of persons working in the field who have asked that the procedures be revised and that the reference-group norms be expanded to cover a wider age range" (Lambert, 1981, p.3). The ABS-SE was designed for use with children and youths ages 3 through 17. The items on the ABS-SE are the same as those on the ABS-PSV.

One of the major revisions to the ABS-PSV is in the interpretive scheme. The ABS-PSV consists of two sections, the first reporting on the adaptive functioning in nine skill areas or domains, the second, on maladaptive behavior in 12 domains. The 21 domain raw scores are converted to percentile ranks and then compared with those in regular and special education classes. It is suggested in the manual (Lambert et al., 1975) that those children who perform similar or worse than 75% of children classified as EMR, for example, on many of the domains, can be comfortably identified EMR.

The ABS-SE contains the same items but provides a refined scheme for interpretation. For diagnostic purposes, a diagnostic profile of the child's performance on five factors is computed. These five factors labeled Personal Self-Sufficiency, Community Self-Sufficiency, Personal-Social Responsibility, Social Adjustment, and Personnal Adjustment were derived from factor analytic studies (Nihira, 1969a; Nihira, 1969b; Guarnaccia, 1976; Lambert & Nicoll, 1976). Factor scaled scores are compared with the norm groups of interest. If the child is being considered for EMR diagnosis, for example, factor scaled scores are compared with EMR and regular class children. Factor scaled scores that are more than one standard deviation below the mean and considered diagnostically significant.

In addition to the factor scaled scores, a Comparison Score composed of the weighted scores on three factors, Personal Self-Sufficiency, Community Self-Sufficiency and Personal-Social Responsibility, is computed to aid in classification and placement. It is suggested that a child performing in the bottom 5 percent of the Regular group signifies the possibility of mental retardation (Lambert, 1981).

The standardization of the ABS-SE is on children and youth in regular EMR, and TMR classes. A total sample of 6,523 children and youth were used from California and Florida. Several studies are reported in the manual offering evidence of the validity of the ABS-SE. Studies on the content, construct, predictive, and outcome validities of the Domain, Factor and Comparison Scores are offered. Several overall conclusions can be reached regarding the validity of the ABS-SE. First, the internal construct validity appears adequate. The ABS-SE is organized around empirically structured clusters of items that were thoughtfully selected and analyzed. Second, adequate evidence is offered concerning the external construct validity of the ABS-SE. Evidence indicates that the various scores derived from Section I of the ABS-SE have, as expected, low to moderate correlations with IQ. Both Section I and II factor scores correlate moderately with standardized tests of achievement. Third, outcome validity of the test is offered to help in the selection of students. Domain scores for both Sections I and II of the ABS-SE discriminate between those placed in regular, EMR, and TMR classes although Section I domain scores appear to discriminate better than Section II scores.

Comparison scores, derived for the most part from items included in Section I factors, show considerable accuracy in identifying the extent to which a child's performance is like students in regular, EMR, or TMR programs.

The internal consistency of the items making up each of the factors is offered as evidence of the reliability of the ABS-SE. Overall, these reliability coefficients are high with three of the factors (i.e. Community Self-Sufficiency, Personal-Social Responsibility and Social Adjustment) sufficiently high to use in interpreting individual profiles. Personal Self-Suffiency and Personal Adjustment reliability coefficients are too low to recommend for individual profile interpretations. In addition to information on the internal consistency of the data, standard error of measurement information is provided in the manual to help in interpretations.

Evidence offered in the ABS-SE manual indicates that neither ethnic status nor sex is associated with performance on the domains in Section I of the ABS-SE. Also, no mean score differences were found on Section I among ethnic classes and between sexes at each of the three levels of classification (i.e. regular, EMR, and TMR) (Cole, 1976). Ethnic status and sex did significantly contribute to Section II performance on the ABS-SE. In the Cole (1976) study of mean differences among ethnic groups within classification showed differences among ethnic groups but the effects of this variable explained only 1 to 2 percent of the variance in performance. Consequently, as a function of there being no substantive mean differences across groups or sexes, there are good chances that the test is unbiased in that it is measuring the same construct equally well for all and will predict equally well for all. However, this is only an inference and empirical evidence should be gathered in its support.

Mastenbrook (1977) criticizes the content of the ABS-PSV which is the same in the ABS-SE in that it emphasizes self-maintenance type behaviors with little regard for behaviors concerning social roles outside the school. This criticism along with the limitations in the standardization of the instrument (i.e.,only two States are represented) need to be taken into account when considering the use of the ABS-SE.

Adaptive Behavior Inventory for Children (ABIC)

The ABIC was developed as part of the comprehensive System of Multicultural Pluralistic Assessment (SOMPA; Mercer, 1979). However, the ABIC can be administered and scored separately. Consistent with the purpose of the SOMPA, the ABIC was designed for the major purpose of classification/placement. The measure has norms for children five through 11 inclusive and in this sense is more limited than the ABS-SE. It employs parents as third party respondents, takes approximately one hour to administer, and can be administered by paraprofessionals with training (see Table 7.7). Its major advantage in comparison to other measures is its assessment of behaviors across a variety of settings. Other measures do not specifically address the child's functioning in different environments as well as the ABIC (see Table 7.7).

The items for the ABIC were derived from a conceptualization of adaptive behavior that conceived it as "an adaptive fit in social systems through the development of interpersonal ties and the acquisition of specific skills required to fulfill the task functions associated with particular roles" (Mercer, 1979, p. 93). The six scales include family, community, peer, nonacademic school, earner/consumer, and self-maintenance. Performance in school, while still conceived as a component of adaptive behavior, is measured by performance on the WISC-R in Mercer's system. The ABIC consists of a total 242 items, reduced from an initial item pool of 480 items. The choice of items was heavily based on intensive interviews with mothers and items chosen for the scales were identified through an analytic sorting procedure.

Special attention was given to the choice of items that were believed to be less likely to show differences across race and sex.

The norms for the ABIC are based on a stratified random sample of 2085 California children between the ages of 5 and 11. The sample was stratified according to ethnic/racial group, sex, age, and size of community. Raw scores derived from the six scales are converted to scaled-scores with a mean of 50 and a standard deviation of 15. Standard error of measurement information is provided so that probability statements can be made regarding the range within which the child's true score lies.

The split-half reliabilities of the various scales, ages, and ethnic/racial groups is provided. They range from .78 to .92 with a median reliability coefficient of .86.

The relationship between the WISC-R and the ABIC as reported are low ranging from near zero to .3 (Kazimour & Reschly, 1980; Mercer, 1979; Tebeleff & Oakland, 1977). Similar correlations between the ABIC and measures of academic achievement are reported (Sapp, Horton, McElroy & Ray, 1979; Tebeleff & Oakland, 1977). In a comparison of scores on the ABIC across racial/ethnic groups (i.e., White, Black and Hispanic), Mercer (1979) reports that the significant differences that were found were too small to be of any practical significance in interpreting scores for individual children. Grindby and Mastenbrook (1977) report scores for lower income Mexican-American children lower than other groups. Sattler (1982) criticizes the ABIC for not providing the opportunity for an assessment of whether or not such differences are a function of decreased opportunities in the child's environment rather than a child's lack of ability. Sattler (1982) provides three additional criticisms of the ABIC. First, it relies exclusively on the questionable responses of parents or guardians. Second, some of the items may discriminate against low SES and minority group children. Third, the norms for the ABIC may not be adequate for use outside of California. Buckley and Oakland (1977), for example, report lower scores for Texas children than California children.

It should be noted that Mercer (1979) argues that the validity of the use of the ABIC should not be based on its relationship to academic achievement or IQ. Rather, the purpose of the ABIC is to assess the extent to which the child is meeting expectations within the social systems he/she is functioning. Consequently, Mercer argues that the predictive utility of the ABIC should be judged against such criteria. However, as Oakland (1979) points out, since these criteria are not available, the predictive validity of the ABIC according to Mercer's definition remains unknown.

Piagetian Assessment Procedures

Although there has been a noticeable lack of reliance on standardized intelligence tests in the development of Piaget's theory (Brainerd, 1978), the clinical methods derived from this area are being used as an alternative to traditional procedures. Piaget conceptualized human development as occurring in a series of invariant stages. In each stage a discrete set of mental operations is purported to be used in organizing experience and adapting to the environment. Although Piaget assumed that the way in which experience was organized is genetically determined, the environment is said to influence the rate of developmental progress (Flavell, 1963).

Piaget outlined four main stages of cognitive or intellectual development. These four main stages are outlined in Table 7.2. The manner in which the child is assessed and the responses scored is usually different from more traditional testing. This perspective is reflected in the statements by Elkind (1974):

> ...when we deal with (the child's thinking processes) we must not evaluate them as right or wrong but rather value them as genuine expressions

Table 7.2

Piaget's stages of cognitive development

Stage	*Approximate age range*	*Primary features, especially toward the end of each stage*
I. Sensorimotor	Birth to 2 years	"Thought" occurs primarily through actions.
		Coordination of sensory input improves.
		Coordination of physical responses improves.
		Objects and people, including self, are differentiated from one another and recognized as permanent.
II. Preoperational	2 to 7 years	Language use and symbolic thought increase.
		Egocentrism predominates.
		Centration (attending to a striking feature or part) rather than decentration (analysis of whole and parts) characterizes perception and thought.
		Produces mental images of static situations and things, rather than of processes and transformations.
		Irreversibility in thought (can think in one way but not its reverse; e.g., counting, saying letters of the alphabet).
		Perceptibly similar objects are classified as alike.
		Words (names) are associated with some things and with some classes of things.
III. Concrete Operations	7 to 11 years	Logical thinking using concrete objects occurs.
		Less egocentric and more socialized speech occurs.
		Conservation increasingly occurs.
		Decentering and reversibility occur.
		Understands changes and processes and more complex static events and relations.
		The same things are grouped correctly into two or more different classes.
		Relations among actual things and classes of things are understood; also relations among words that represent things and classes of things that have been experienced are understood.
IV. Formal Operations	11 years to adult	Mental operations in symbolic form are carried out and operations are performed on ideas as things.
		Comparisons, contrasts, deductions and inferences from ideational content rather than concrete things and events.
		Relations between and among symbols standing for concepts that have not been experienced directly are understood.

Source: Adapted from Ginsburg and Opper, 1969.

Source: Klausmeier, H.J., & Goodwin, W. Learning and human abilities: Educational psychology. New York: Harper & Row, 1971. (adapted from Ginsburg and Opper, 1969).

BEST COPY AVAILABLE

> of the child's budding mental abilities. When we deal with spatial, temporal, causal, or quantitative concepts, we need to explore the kinds of meanings children give to such terms. Such exploration reveals the level and reference frame of the child's understanding. More importantly, such exploration avoids the inhibiting suggestion that the child's incomplete (but partially correct) understanding of such terms is "wrong ". A teacher who sees a child's productions as having value, as meaning something, avoids putting the child on the tract of always seeking "right" answers. More importantly, perhaps, her orientation conveys to the child a sense of her attempt to understand him and her respect for her intellectual productions (p. 125).

An important aspect of Piagetian assessment is that the course of cognitive development is said to be invariant in sequence. In this regard a child can be in one of the four stages (or in a transitional stage) in which he/she performs tasks within a given stage. The child cannot miss a stage in the usual sense because various cognitive structures or schemes serve as the basis for all normal development. Thus, the sequential mental development is said to occur in all children regardless of race or social class. In this regard, the Piagetian tasks may be less culture loaded than conventional IQ measures (Jensen, 1980).

Some writers have noted that the Piagetian tasks are less susceptible to influence by specific instruction than the usual IQ measures (e.g., Kohlberg, 1968; Sigel & Olmsted, 1970). Some contrasts between the Piagetian and more traditional psychometric approaches to intelligence are presented in Table 7.3. As noted by Elkind (1974), the two approaches differ on the following dimensions: The type of genetic causality which they presuppose, the conceptions of the course of mental development, and the relative contributions of nature and nurture to intellectual skills.

There have been relatively few Piagetian assessment measures developed for use in applied settings. However, Struthers and DeAvila (1967) developed the Cartoon Conservation Scales which is a test for children that can be administered on a group basis. The test seems to be appropriate as a measure of cognitive development with respect to certain aspects of the Piagetian conservation concept. This assessment approach may prove valuable in that there appears to be a similarity in cognitive development of children from diverse cultural background when assessed on certain Piagetian tasks (DeAvila & Harassy, 1975).

A number of Piagetian assessment procedures are reviewed by Johnson (1976). However, many of these tap specific skills (e.g., conservation of number, Swanson, 1976a) and represent

Table 7.3

Comparison of Piagetian and Psychometric Approaches to Intelligence

Similarities	*Differences*	
	Piagetian	*Psychometric*
1. Both accept genetic determinants of intelligence. 2. Both accept maturational determination of intelligence. 3. Both use nonexperimental methodology. 4. Both attempt to measure intellectual functions that the child is expected to have developed by a certain age.	1. Assumes that there are factors which give development a definite, nonrandom direction. Mental growth is qualitative and presupposes significant differences in the thinking of younger versus older children; concerned with intra-individual changes occurring in the course of development.	1. Tested intelligence is assumed to be randomly distributed in a given population, with the distribution following the normal curve; concerned with inter-individual differences.
5. Both conceive of intelligence as being essentially rational. 6. Both assume that maturation of intellectual processes is complete somewhere during late adolescence.	2. Views mental growth as the formation of new mental structures and the emergence of new mental abilities.	2. Views the course of mental growth as a curve which measures the amount of intelligence at some criterion age that can be predicted from any preceding age.
7. Both are capable of predicting intellectual behavior outside of the test situation.	3. Genetic and environmental factors interact in a functional and dynamic manner with respect to their regulatory control over mental activity.	3. Genetic and environmental contributions to intelligence can be measured.

Note. Similarity items 5, 6, and 7 obtained from Dudek, Lester, Goldberg, and Dyer (1969); the remainder of the table adapted from Elkind (1974)

Source: Sattler, J.M. Assessment of children's intelligence and special abilities (2nd. ed.) Boston: Allyn and Bacon, 1982.

BEST COPY AVAILABLE

"experiemental" or "research" instruments at this time. Thus, their usefulness in nonbiased assessment remains unknown. However, within the context of a broadened scope of assessment, these various devices may be useful in the assessment of certain specific skills (Johnson, 1976). One commercially produced set of Piagetian assessment procedures is the Concept Assessment Kit (CAK) (Goldschmidt & Bentler, 1968a). The CAK is for use in individual assessment of children in such areas as conservation of number, substance, weight, two-dimensional space, and continuous and discontinuous qualities. The test has been reviewed in Buros 7:437, the Journal of Educational Measurement, 1969, 6, 263-269, and briefly by Jensen (1980). Generally, the CAK has somewhat limited norms (i.e., 560 children in the Los Angeles area) and a somewhat limited age range.

One of the more extensive reviews of Piagetian assessment and the issues surrounding these measures in test bias was presented by Jensen (1980). He noted that Piagetian procedures show promise as culture reduced measures; but he raised two important questions regarding these measures: "Do Piaget's tests measure a different mental ability than the *g* measured by conventional IQ tests? and (2) do minority children and (culturally disadvantaged) children perform better on the Piagetian tests, relative to majority children, than on conventional IQ tests?" (p. 673). In response to the first question Jensen (1980) notes that the correlations between various intelligence and achievement tests and various

Piagetian measures of from 5 to 10 scales assessing concrete operations range from .18 to .84 (X=.50; see Table 7.4). In the case of the Garfinkle (1975) study, Jensen (1980) conducted a principal components analysis of the intercorrelations among the 14 Piagetian tasks. He found that the squared multiple correlation of each item with every other item is comparable to that found on the Wechsler subtests.

Jensen (1980) also reports that Piagetian tests show social class and ethnic group differences in the United States with children from low SES backgrounds about as far behind in the Piagetian measure as for more traditional IQ tests (e.g., Almy, 1970; Almy, Chittenden, & Miller, 1966; Figurelli & Keller, 1972; Tuddenham, 1970; Wasik & Wasik, 1971). Jensen (1980) notes:

> In all such comparisons of group measures, one must take into account the small number of items of the Piagetian tests, which tends greatly to attenuate mean differences expressed in units or standard score units. When this is properly taken into account, in terms of *item* discriminabilities and inter-item correlations, it turns out that the Piagetian tests show *larger* white-black differences than the Stanford-Binet or other conventional IQ tests. I figure the white-black mean difference in

Table 7.4

Table 7.4 Correlation (*r*) between Piagetian tests and various measurements of intelligence and scholastic achievement.

Variable	*r*	Study
Intelligence Tests		
Stanford-Binet MA	.38	Beard, 1960
WISC MA	.69	Kuhn, 1976
WISC Full Scale IQ	.43	Elkind, 1961
WISC Verbal IQ	.47	Eklind, 1961
WISC	.69-.84	Hathaway, 1972
Raven's Matrices	.60	Tuddenham, 1970
Peabody Picture Vocabulary Test	.21	Tuddenham, 1970
Peabody Picture Vocabulary Test	.47	Gaudia, 1972
Peabody Picture Vocabulary Test	.28	De Avila & Havassy, 1974
Peabody Picture Vocabulary Test	.31	Klippel, 1975
Lorge-Thorndike MA	.62	Kaufman, 1970, 1971
Lorge-Thorndike IQ	.55	Kaufman, 1970, 1971
Gesell School Readiness Test	.64	Kaufmann, 1970, 1971
IQ-Unspecified Test	.24-.34	Dodwell, 1962
Mean r	.49	
Scholastic Achievement		
Reading (SAT)	.58	Kaufmann & Kaufmann, 1972
Reading (SAT)	.42	Garfinkle, 1975
Arithmetic (SAT)	.50	Garfinkle, 1975
Arithmetic (SAT)	.60	Kaufman & Kaufman, 1972
Mathematics (MAT)	.18-.41	De Vries, 1974
Arithmetic Grades	.52	Goldschmidt, 1967
Composite Achievement (SAT)	.64	Kaufman & Kaufman, 1972
Composite Achievement (California Achievement Test)	.63	Dudek et al., 1969
Mean r	.55	

Source: Jensen, A.R. Bias in mental testing. New York: The Free Press, 1980.

> units would be about 20 percent larger than the Stanford-Binet IQ difference on Piagetian tests of comparable length to the Stanford-Binet. But, while Piagetian tests tend to magnify the white-black difference, they tend to diminish the differences between whites and Mexicans and Indians, and Orientals tend to surpass whites in Piagetian performance. Interestingly, Artic Eskimos surpass white urban Canadian children on Piagetian tests, and Canadian Indians do almost as well as Eskimos (p. 676).

Learning-Potential Assessment

Generally, the learning-potential approach views assessment as an examination of learning and strategies which facilitate acquisition of new information or skills (cf. Kratochwill, 1977). Learning-potential assessment bears similarity to Piaget's work on intellectual development (Haywood, Filler, Shifman, & Chatelant, 1974). That is, within the Piagetian paradigm, intelligence is viewed as a process rather than a static entity unmodifiable by experience.

Work in the learning potential area has been affiliated with Haywood and his associates in Nashville, Tennessee; Budoff in Cambridge, Massachusetts; and Feurstein and his associates in Jerusalem. These investigators and their collegaues have adapted test-based models for assessment and intervention of the mentally retarded and/or learning disabled (Haywood et al., 1974 and Kratochwill, 1977, for overviews). Haywood et al. (1974) noted that verbal abstraction abilities can be improved in mental retardation associated with culturally different environments during actual assessment. For example, some research indicates that mentally retarded clients are able to perform better on Wechsler's Similarities subtest when examples of each concept are provided (Gordon & Haywood, 1969). These results apparently replicate with retarded children and adults from culturally different environments (Haywood & Switzky, 1974).

The learning-potential work of Budoff and his associates has used a test-train-retest assessment paradigm on such instruments as the Kohs' Block Design Test (Budoff, 1967), Wechsler Performance Scale (Budoff, 1969), Raven's Progressive Matrices (Budoff & Hutton, 1972), and a modification of Feurstein's (1968) early Learning Potential Assessment Device (Budoff, 1969). These tasks are sensitive to modification via instruction or coaching and typically assessment can yield three types of performers. High-scorers gain little from coaching. Those who initially score low and demonstrate performance gains following instruction are labeled gainers. Nongainers initially score low but do not show gains following training (see Budoff, Meskin, & Harrison, 1971). A major implication of Budoff's work has been that:

> A large proportion of IQ-defined retardates, who

> come from low income homes and have no history of brain injury, show marked ability to solve these tasks when they are presented in the learning potential assessment format. The data indicate that the more able students by this criterion are educationally, not mentally, retarded, and the ability they demonstrate prior to or following tuition is not specific to the particular learning potential task (Budoff, 1972, p. 203).

There is some empirical work on the learning potential strategy. For example, Budoff and Hutton (1972) found that if they provided only an hour of structured experiences in problem solving to children who initially scored low on the Raven's, 50 percent of these low performance children scored at the 50th percentile (or above) on a posttest administered after training. These gainers represented minority groups. Similar results have been found with "learning disabled children' (e.g., Platt, 1976; Swanson, 1976). Sewell and Severson (1974) also found that the Raven Progressive Matrices (see Budoff & Friedman, 1964) usefully differentiated low SES black children who could profit from learning experiences. Nevertheless, unanswered questions in this area relate to how learning potential assessment yields prescriptive information for classroom instruction, especially in various academic content areas (math, reading, etc.) and the generalizability of the training (cf. Kratochwill, 1977).

Another area within the learning potential paradigm is represented in the work of Feurstein and his associates (Feuerstein, 1968, 1970; Feuerstein & Rand, 1978). Like the test-train-test paradigm of Budoff and his associates, Feuerstein's strategy is designed to promote the best possible learning and motivational conditions of the child. The Learning Potential Assessment Device (LPAD) is designed to assess what an individual can learn rather than the traditional inventory of what one has learned and current problem-solving ability (Feuerstein & Rand, 1978).

A detailed discussion of the LPAD can be found in Feuerstein, Rand, and Hoffman (1979). The LPAD is usually employed for individual assessment, but a group version has been developed. In the group test students are assessed on tasks that become progressively more difficult. The conceptual framework for the LPAD is as follows:

> The assessment of learning potential differs from that of standardized psychomtric techniques in a number of significant ways. The primary differencee lies in the conceptual foundations upon which the assessment is based. In place of the static goals generated by conventional psychometric theory and techniques which determine the nature and structure of its measuring instruments, the LPAD and its theoretical framework, the cognitive map, generate dynamic goals which reflect the underlying dimensions of the adaptive processes

involved in intelligent activities. In terms of the actual techniques employed, the central purpose is again very different. Tests that yield IQ measures are constructed to provide a reflection of an individual's manifest level of performance relative to other individuals within a representative, normally distributed population. The LPAD is geared toward producing changes within the individual during the testing situation in order to permit an ongoing assessment of that individual's ability to learn and change relative to his/her own optimal levels (Feuerstein, Miller, Rand, & Jensen, 1981, pp. 202-203).

Feuerstein et al. (1980) note the learning potential assessment requires four specific conceptual shifts from traditional testing.

1. A shift from product to process orientation. In this regard, the LPAD is designed to alter the individual's performance during the actual assessment.
2. The test structure includes the conceptual features of the cognitive map. Figure 7.1 shows the structural model of the LPAD and Figure 7.2 presents an example of the test instrument. The task is

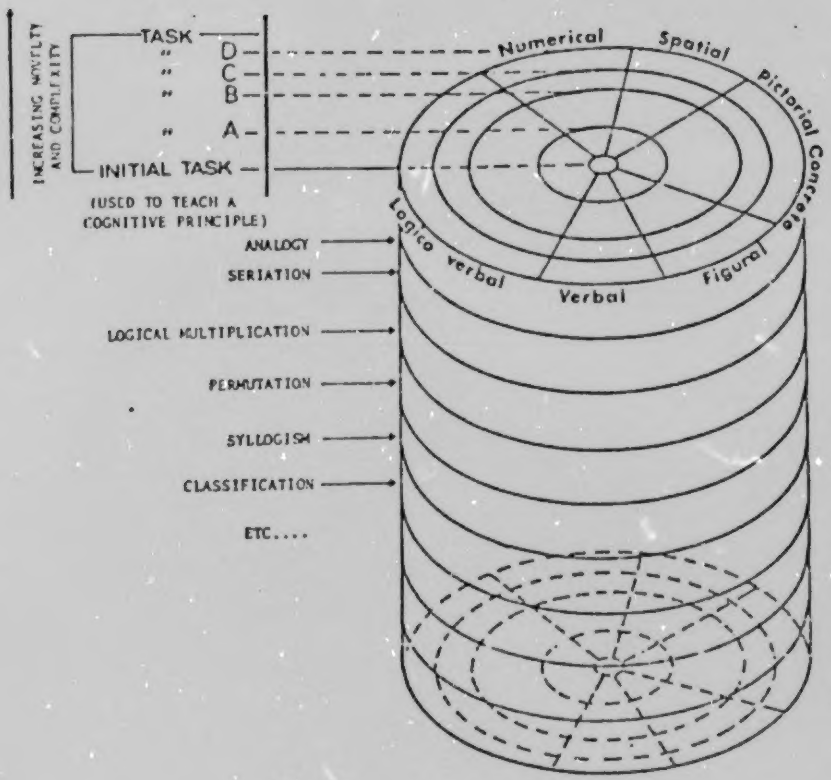
INCREASING NOVELTY AND COMPLEXITY
TASK
" D
" C
" B
" A
INITIAL TASK
(USED TO TEACH A COGNITIVE PRINCIPLE)
Numerical
Spatial
Pictorial
Concrete
Figural
Verbal
Logico verbal
ANALOGY
SERIATION
LOGICAL MULTIPLICATION
PERMUTATION
SYLLOGISM
CLASSIFICATION
ETC....

Figure 7.1 - The Learning Potential Assessment Device Model (Source: Feuerstein, R., Miller, R., Rand, Y., & Jensen, M.R. Can evolving techniques better measure cognitive change? The Journal of Special Education, 1981, 15, 201-219. Reproduced by permission).

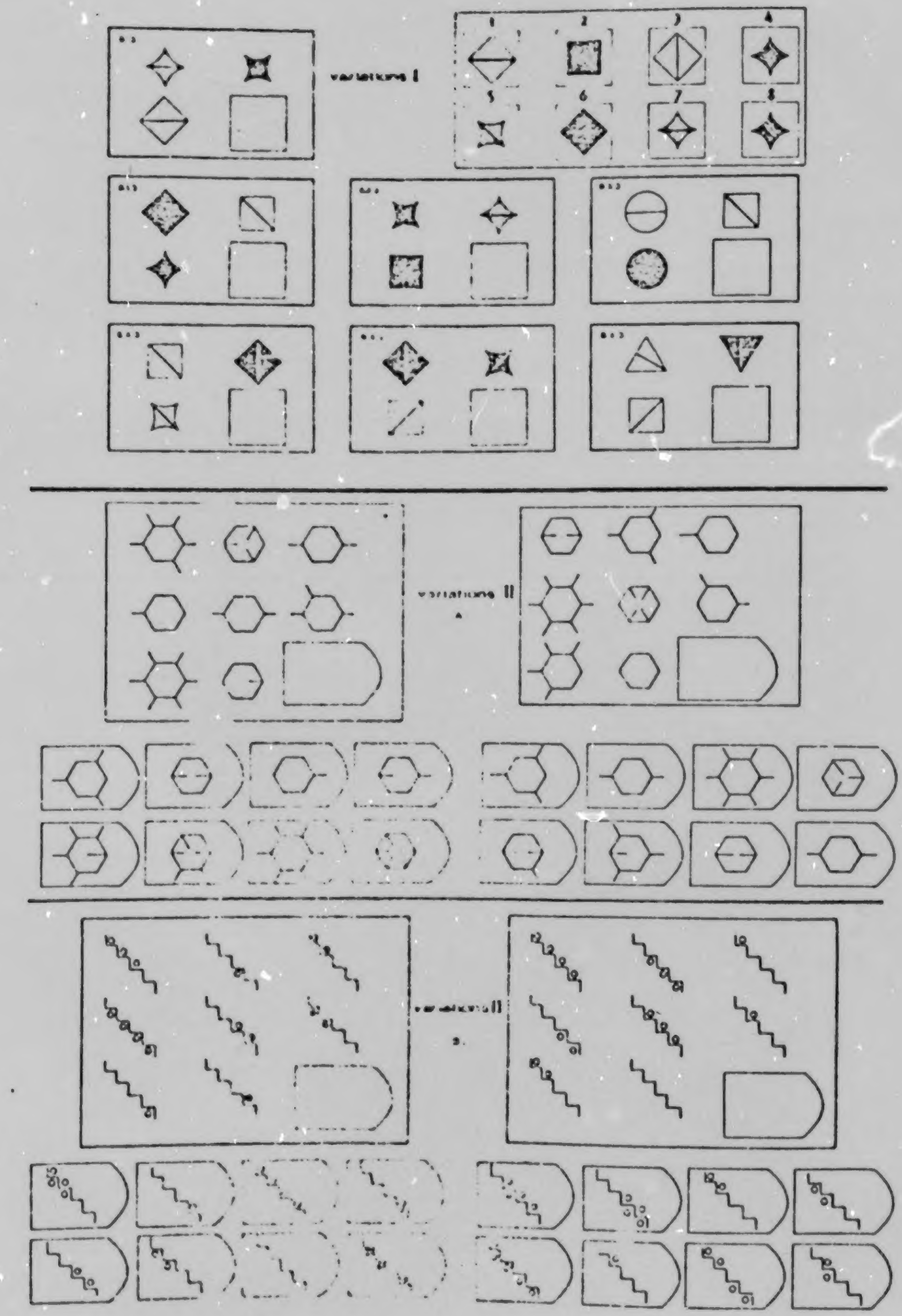
variations I
1
2
3
4
5
6
7
8
variations II
variations II

Figure 7.2 - Example of Raven's Matrices variations based on LPAD Model (Source: Feuerstein, R., Miller, R., Rand, Y., & Jensen, M.R. Can evolving techniques better measure cognitive change? The Journal of Special Education, 1981, 15, 201-219. Reproduced by permission).

presented in a variety of modalities (e.g., numerical, verbal, figural) and could require a number of different operations (e.g., analogy, classification, and seriation). The student is given training to solve a particular problem and it is assumed that training provide a temporary correction of deficient prerequistie functions. Data like the following are obtained.

a. The capacity of the examinee to grasp the principle underlying the initial problem and to solve it;

b. The amount and nature of investment required to teach the examinee the given principle;

c. The extent to which the newly acquired principle is successfully applied in solving problems that become progressively more different from the initial task;

d. The differential preference of the examinee for one or another of the various modalities of presentation of a given problem; and

e. The differential effects of different training strategies offered to the examinee in the remediation of his/her functioning; these effects are measured by using the criteria of novelty-complexity, language of presentation, and types of operation (Feuerstein et al., 1979).

3. The test situation is changed with one of a test-teach-test procedures. Factors that produce change, including actively instructing the student, feedback concerning success and failure, engaging the child in the learning process, promoting intrinsic motivation, interpreting performance, and individualizing the test situation, are used during the actual assessment.
4. The results of the LPAD are interpreted differently. For example, considerable weight is given to excellent responses, and unexpected responses are pursued (i.e., a process approach is taken).

More recently, Arbitman-Smith and Haywood (1980) described an educational program that is used to enhance the learning skills of slow learners and learning disabled students. The approach, developed by Feuerstein and his colleagues (Feuerstein, 1979; Feuerstein, Rand, Hoffman, & Miller, 1980) is called Instrumental Enrichment (IE). Arbitman-Smith and Haywood (1980) describe the IE program:

> It not only encourages students to engage in a learning process that is different from their past experiences and thus not associated with past failures, but also gives teachers an organized, structural framework with which to teach a variety of problem-solving processes. The program emphasizes processes by which various problems are solved rather than merely getting "right" answers.

> Basic cognitive operations such as evaluation, interpretation, planning, and comparison are consistently taught through the program,and the students are made aware of their own thought processes. This consists currently of 15 teaching instruments, each focused upon (but not limited to) a specific deficient cognitive function. The program constitutes in the aggregate approximately 250 to 300 hours of classroom instruction, 3 to 5 hours a week, to be spread over at least two years. (p. 53).

More detailed information on the content and goals of the IE program can be found in Feuerstein et al. (1980).

The IE program has been evaluated in Israel (Feuerstein, Rand, Hoffman, & Miller, 1980) and in the United States and Canada (Haywood & Arbitman-Smith, 1980). Arbitman-Smith and Haywood (1980) provided an overview of some preliminary data on cognitive education of learning disabled students. The authors reported that data have been collected on two one-year contracts. They note that there were generally no first year effects on school achievement, but some changes in intellectual functioning were found (see Haywood & Arbitman-Smith, 1980, for an overview). With their LD population Arbitman-Smith and Haywood (1980) report that students "indicated interest and motivation to learn the IE program materials and actively participated in the discussions, a form of behavior not often exhibited in their regular classes" (p. 62). Unfortunately, even these effects may not have been due to the program since no control for such effects was included in the study.

Some writers have noted that the work in the learning potential assessment area appears promising for the nondiscriminatory or nonbiased assessment (Alley & Foster, 1978; Laosa, 1977; Mercer & Ysseldyke, 1977). Indeed, Mercer and Ysseldyke (1977) include the learning potential assessment paradigm as part of the pluralistic assessment "model." Feuerstein et al. (1979) have noted that the LPAS provides a more fair assessment of minority students than attempts to adapt conventional psychometric tests. They note that such techniques as culture-free, culture fair, developmental tests, and the SOMPA procedures perpetuate a confusion between manifest performance and potential (p. 203).

There is some empirical research on the LPAD (see Feuerstein et al., 1979, for a detailed review). Research has been conducted with disadvantaged children, on homogeneous versus heterogeneous grouping, and on assessment of culturally different immigrants. In the latter group Feuerstein et al. (1981) report that minimum training provided by the group test procedure produced substantial learning potential and higher levels of cognitive modifiability. More traditional measures apparently reflected cultural differences and not differences in ability to learn or profit from instruction.

Although the LPAD and the IE program appear somewhat promising, there is still little evidence to suggest that these measures are any less biased than traditional measures. Their merit presumably lies in teaching during the actual testing session, but it is not known if the cognitive strategies that are trained have any relation to the child's performance in the classroom setting.

Diagnostic/Clinical Teaching

An area that bears similarity to the learning potential strategies is called "diagnostic clinical teaching" (Kratochwill, 1977; Lerner, 1976). These strategies differ from "diagnotic-prescriptive teaching" (Ysseldyke, 1973; Ysseldyke & Salvia, 1974; Salvia & Ysseldyke, 1978) which have been affiliated with test-based aptitute-treatment interaction (ATI) paradigms (e.g. Cronbach & Snow, 1976; Levin, 1977). Diagnostic teaching actually embraces a number of different strategies which are, at present, not guided by any particular theoretical area. Typically, diagnostic teaching involves the actual teaching of curriculum-related material under conditions that maximize learning (e.g., stimulus materials, mediational strategies, reinforcement, feedback). Thus, their relevance in the nondiscriminatory assessment area is that they focus on tasks nearly all children experience in the school curriculum and they focus on direct intervention for successful curriculum mastery. For example, Myers and Hammill (1969, 1973) recommended teaching words to children under conditions that maximize learning and suggested that learning disabled children should be evaluated on learning tasks from which norms are established.

Likewise, Hutson and Niles (1974) proposed trial teaching as a supplement to traditional testing. Severson (1971, 1973) suggested a process learning assessment strategy based on teaching academic content under different conditions. In research tasks employing from four to eight words to be learned, predictive validity relations have ranged from .30 to .73 with achievement test criteria (see Kratochwill & Severson, 1977; Sewell & Severson, 1974).

More recently Sewell (1979) examined the predictive effectiveness of intelligence tests and learning tasks for first grade black and white children. The study focused on the relative merits of learning tasks in contrast to traditional IQ tests in predicting academic achievement. The learning tasks involved diagnostic teaching, paired associate learning, and a learning potential assessment using the Raven's Coloured Progressive Matrices in a pretest-coaching posttesting format. The diagnostic teaching condition involved teaching the children 15 words under three different conditions that proceeded from feedback to social praise to tangible reinforcement. The Stanford-Binet served as the measure of intelligence and the California Achievement Test as the criterion. The results indicated that the IQ measure correlated moderately with achievement with both groups, the

IQ was a more reliable predictor for the white children than for the blacks, and for the black children, certain learning tasks were better predictors than IQ. The study does show that although the IQ was a significant predictor for both groups, both this measure and the learning tasks were a better predictor of achievement for the middle-class children.

While these procedures represent a promising area within nondiscriminatory assessment, a paucity of research and a limited range of content remain limitations (cf. Kratochwill, 1977). The diagnostic teaching procedures are also said to represent information useful for prescriptions (e.g., Sewell, 1979, 1981). Unfortunately, it is not at all clear how specific educational programs are to be developed from the diagnostic teaching procedure. Presumably, information on how the child learns academic material under various conditions of reinforcement can be obtained. However to date, the research on diagnostic teaching has involved a rather limited set of dimensions that are known to influence the learning process. It is doubtful that assessment under the usual conditions of diagnostic teaching will generalize to the child's learning in other settings such as the classroom.

Child Development Observation

Within a tradition similar to the learning potential assessment and diagnostic teaching is the "Child Development Observation" (CDO) designed by Ozer and his associates (Ozer, 1966, 1968, 1978; Ozer & Dworkin, 1974; Ozer & Richardson, 1972, 1974). A major objective of CDO is to simulate the process of learning on protocols that sample conditions under which a given child's learning problems may be solved. Different teaching strategies are also enacted to see how the child best learns.

The CDO procedure may be useful in nonbiased assessment in that it does not conform to traditional testing paradigms; no score is derived in relation to a norm group; decisions do not promote diagnostic labeling; and relating assessment data to classroom functioning is intrinsic to evaluation (Ozer & Richardson, 1972). However, there are no data on the reliability and validity of the procedure; verbal skills are heavily emphasized in certain areas of assessment, and the CDO does not systematically sample from classroom tasks (cf. Kratochwill, 1977). Like the other process oriented measures of learning potential and diagnostic teaching, the CDO may not reflect the conditions under which learning usually occurs in the child's usual educational environment.

Clinical Neuropsychological Assessment

The field of neuropsychology is concerned with delineating brain-behavior relations. Neuropsychology includes a number of different, sometimes only remotely related, disciplines of which clinical neuropsychology is but one. Clinical neuropsychology focuses on developing knowledge about human brain-behavior relations, or delineating the psychological correlates of brain lesions (Davison, 1974; Reitan, 1966). Intellectual, sensory-motor, and personality deficits are measured and related to brain lesions or to brain damage in the broader sense of physiological impairment. The work in this area is rooted in academic psychology, behavioral neurology, and particularly in the psychometric field in psychology.

Within clinical neuropsychology there is a dependence upon standardized behavioral observations emphasizing normative psychological assessment devices. Within this context, behavior is defined operationally and, usually, quantified along continuous distributions (Davison, 1974). The clinical neuropsychologist is typically not merely concerned with distinguishing brain damage from other conditions. Rather, the interest lies in refining descriptions of clinical conditions including inferences relative to location and extent of brain damage, as well as probable medical and psychological conditions accounting for the abnormal behavior.

A considerable amount of information has been obtained during the past decade about the behavioral characteristics of brain-damaged persons as a result of neuropsychological study in the areas of mental retardation, learning disabilities, behavioral disabilities, and convulsive disorders (cf. Reitan & Davison, 1974). In addition, studies have been conducted on individuals with confirmed cerebral lesions independently of whether these individuals manifest learning or behavioral problems (Reitan, 1974). Finally, neuropsychological studies of normal children have been undertaken (e.g., Kimura, 1967).

Further research in neuropsychological assessment techniques could lead to some interesting applications relative to identification, classification, and intervention strategies. The concept of brain dysfunction as a primary factor in learning disabilities (LD), for example, has received increasing attention over the past 20 years. By characterizing all children having learning disabilities as having minimal brain dysfunction, many professionals seem to have attributed LD to neurogenic factors. However, much of the research relevant to this hypothesized relationship is clouded by the problem that LD children do not constitute a homogeneous group (see, for example, Hallahan & Kauffman, 1978).

Typically, the term "learning disabilities" has been used to refer to children who show a discrepancy between current levels of school performance and measures of academic potential which is not due to mental retardation, cultural, sensory, or educational inadequacies, or serious behavioral disturbances (cf. Bateman & Schiefelbusch, 1969). This type of general definition lacks sufficient objective criteria, so that children who have specific disabilities in reading, spelling, arithmetic or multiple deficits are all categrozied as LD children. Moreover, each type has often been referred to using the term "minimal brain dysfunction." The lack of precision with which professionals have used the terms minimal brain dysfunction and learning disability may partially account for the inconsistencies found in identification and placement practices.

Recent studies in neuropsychological assessment techniques, such as one conducted by Ahn (1977), offer the promising possibility for development of a multiple discriminate function utilizing relevant information for more precise classification of large groups of learning disabled children. Ahn (1977) found significant patterns of difference between three different groups of presumably learning disabled children (i.e., verbal underachievers, arithmetic underachievers, and mixed underachievers) and normal children in quantitative electrophysiological measures (i.e., electroencephalographic evoked potentials).

Results such as these lend plausibility to the contention that neuropsychological assessment techniques may prove useful for more accurate identification and classification of children possessing different specific learning disabilities. At the very least, further research in this area should increase educators' and psychologists; knowledge of the many different types of problems referred to under the general label of "learning disabilities."

Davison (1974) has discussed the potential utility of clinical neuropsychological assessment techniques relative to intervention. Of particular import here is the fact that the same behavioral deficits may be due to differing causal factors and, therefore, require very different interventions. A reading problem for example, may be due to an abnormal learning history or to a structural abnormality of the brain. Thus, for remedial purposes, the etiology of a particular deficit may take on importance. One problem with traditional methods of psychodiagnostic assessment is that they are not typically able to differentiate among the many possible etiological factors involved in a particular disability.

One cannot accurately predict the outcomes of further investigation into this area as yet. Increasing our understanding of brain-behavior relationships will require extensive study of the behavior of humans with brain damage of varying location, extent, etiology, etc. It may be that the product will be merely some interesting descriptive statistics. Undoubtedly, however, increased knowledge of brain behavior correlates holds potential implications for nondiscriminatory assessment techniques as well as decisions based on assessment data. Much additional work needs to be conducted with children, making these procedures more applicable to the area of nonbiased assessment.

Behavioral Assessment Strategies[6]

We have already provided a relatively thorough discussion of the differences between behavior assessment and traditional assessment (see Chapter 3). Nevertheless, the reader is referred to some excellent overviews of this issue (e.g., Ciminero, 1977; Goldfried, 1976; Goldfried & Kent, 1972; Goldfried & Linehan, 1977; Goldfried & Sprafkin, 1974; Mash & Terdall, 1981; Kanfer & Saslow, 1969; Mischel, 1968). Behavioral approaches emphasize a careful examination of environmental antecedents and consequents, as related to a specific response repertoire. Essentially, such an analysis is based on the operant tradition (Bijou & Grimm, 1975; Bijou & Peterson, 1971; Browning & Stover, 1971; Gelfand & Hartmann, 1975). However, as Evans and Nelson (1977) have observed, the strict operant approach can be unduly restrictive. A more global approach, given the current practices in psychology and education, is to outline how a functional analysis approach can utilize more traditional psychometric practices, a rapprochement between traditional psychometrics and social learning theory called "social behavioral psychometrics" by Staats (1975). Moreover, a cognitive functional approach as outlined by Meichenbaum (1977) seems especially useful in some areas of psychological assessment. In comparing this approach to a more conventional functional analysis, Meichenbaum (1977) suggests:

> A cognitive-functional approach to psychological deficits is in the same tradition but includes and emphasizes the role of the client's cognitions (i.e., self statements and images) in the behavioral repertoire. In short, a functional analysis of the client's thinking processes and a

> careful inventory of his cognitive strategies are conducted in order to determine which cognitions (or the failure to produce which key cognitions), under what circumstances, are contributing to or interfering with adequate performance (p. 236).

In a later section of this chapter the cognitive functional analysis strategy will be described in more detail. This strategy as well as others should be conceptualized as expanding the assessment base of psychoeducational behavior assessment.

Conceptual Framework for Behavioral Assessment

With increasing diversity in behavioral assessment, a conceptual framework for classifying behavioral measures is helpful to organize methods and what they are designed to assess. Cone (1977, 1978) and Cone and Hawkins (1977) developed a conceptual framework and a taxonomy called the Behavioral Assessment Grid (BAG). It is based on the simultaneous consideration of three aspects of the behavioral assessment process: (a) the contents assessed, (b) the methods used to assess them, and (c) the types of generalizability (i.e., reliability and validity) established for the measure being employed. The relations among these three aspects of assessment are presented in Figure 7.3.

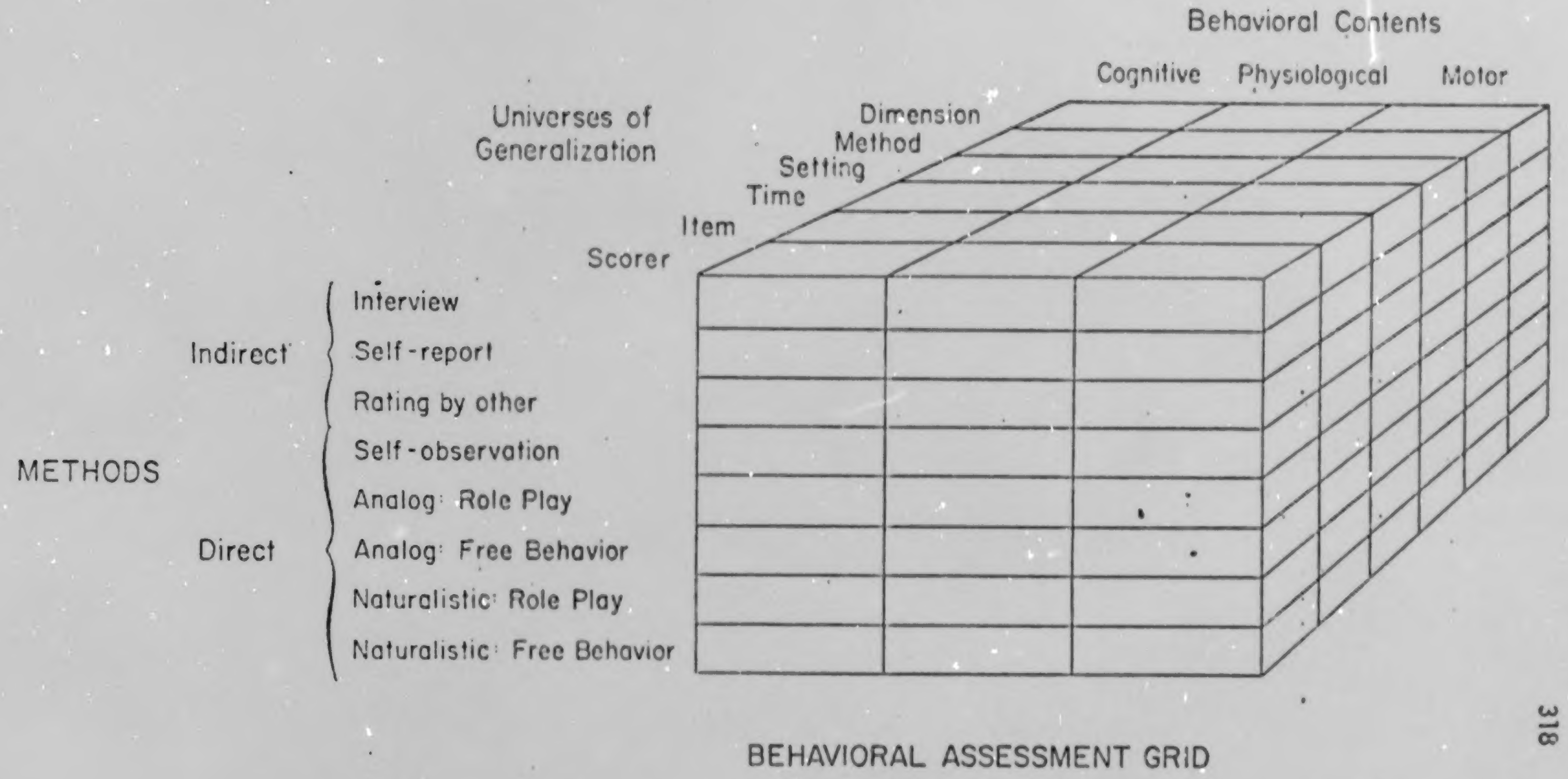
Behavioral Contents
Cognitive
Physiological
Motor
Universes of Generalization
Dimension
Method
Setting
Time
Item
Scorer
METHODS
Indirect
Interview
Self-report
Rating by other
Direct
Self-observation
Analog: Role Play
Analog: Free Behavior
Naturalistic: Role Play
Naturalistic: Free Behavior
BEHAVIORAL ASSESSMENT GRID

Figure 7.3 - The Behavioral Assessment Grid (BAG), a taxonomy of behavioral assessment integrating contents, methods, and universes of generalization (Source: Cone, J.D., The Behavioral Assessment Grid (BAG): A conceptual framework and a taxonomy. *Behavior Therapy*, 1978, 9, 882-888. Copyright 1978 by Association for Advancement of Behavior Therapy. Reproduced by permission).

Contents. Behavioral assessment is commonly conceptualized in three content areas (Cone, 1977; Cone & Hawkins, 1977), systems (Lang, 1968, 1979, 1977) or channels (Paul & Bernstein, 1973). The contents are most commonly referred to as motor, physiological, and cognitive (see Morris & Kratochwill, 1983). Although there are some basic disagreements as to what is specifically included within these categories, we will present the scheme developed by Cone (1978) because it serves a heuristic function.

Motor content is one of the most frequently used content areas and includes activities of the striate musculature typically observable without special instrumentation. Included in this content area would be such activities as walking, running, jumping, talking and other motor components.

Physiological contents, according to Lang (1971), include activities of muscles and glands autonomically innervated and tonic muscle activity. Some examples of physiological content are muscle tension, heart rate, respiration, and galvanic skin response. Such measures are usually assessed through special instrumentation.

Cognitive contents are defined in the context of the particular referents used. Thus, while verbal behavior (self-report) can be categorized as motoric when one is referring to the speech act (see above), the referents may be motor, cognitive, or physiological. When verbal behavior refers to private events (e.g., feelings, thoughts) the referents are cognitive, but when it refers to a publically verifiable behavior, the referent is either physiological or motor. When conducting behavioral assessment, the assessor should be concerned about the relations among the three content areas. This means that in a particular situation an individual may respond cognitively, motorically, and/or physiologically. Some evidence suggests that the three content areas are not necessarily highly intercorrelated (see Bellack & Hersen, 1977a, 1977b; Cone, 1976a; Hersen & Bellack, 1978), but reasons for this remain somewhat unclear (Hugdahl, 1981; Kozack & Miller, 1982). Part of the difficulty in research investigating the relations among the three systems may be related to methodological problems. For example, Cone and Hawkins (1977) argued that comparisons of the three systems have confounded method of assessment with behavioral content. This problem occurs when self-report measures of cognitive activities are compared to direct observation measures of behavior. A child may be trembling but may report that he/she is not frightened. This could be assessed through self-report measures and direct observation, but a low correlation between content areas may be due to content differences or method differences, or both.

A second problem in this area is related to definitions of the three response systems (Hugdahl, 1981; Kozack & Miller, 1982). Such individuals as Lang and Paul and Bernstein have based their definitions on hypothetical constructs. For example, when Lang (1968, 1971) discusses the three response systems in the context of measuring fear, the response is presumed to underly a variety of behaviors such as escape and avoidance. In contrast to this view, Cone and his associates (Cone, 1975, 1976b, 1978; Cone & Hawkins, 1977) prefer the conceptualization in which each content area is examined within the context of stimulus and consequent variables present in any given situation. This latter strategy seems most useful in advancing work in this area, although this still remains debatable.

Methods. Different methods are used to gather data across each of the three content areas. Cone (1977, 1978) ordered these assessment methods along a continuum of directness representing the extent to which they (a) measure the target behavior of clinical relevance and (b) measure the target behavior at the time and place of its natural occurrence.

The methods are categorized into direct and indirect dimensions. Interviews and self-reports are at the indirect end of the continuum because the behavior is considered a verbal representation of more clinically relevant activities taking place at some other time and place. Moreover, ratings by others are included in the indirect category because they typically involve retrospective descriptions of behavior. In contrast to direct observation, a rating of a behavior occurs subsequent to the actual occurrence of the behavior.

Included within the direct assessment methods are self monitoring, analog:role play, analog:free behavior, naturalistic:role play, and naturalistic:free behavior. These dimensions are organized according to who does the observing, the instructions given, the observer, and where the observations occur. In self-monitoring the observer and the observee are the same individual. Self-monitoring differs from self-report in that an observation of the behavior occurs at the time of its natural occurrence. Analogue assessment refers to settings or situations that are analogous to, but not the same as the natural environment. In this type of assessment, the client may be instructed to role play a particular behavior or act normally, as if he/she were in the natural environment. Technically, analogue assessment can vary along a number of dimensions (Kazdin, 1980). Finally, assessment may be scheduled in the natural environment under either role play or completely naturalistic conditions. Each of the eight assessment methods are discussed in more detail in this chapter.

Universes of generalization. The various measures are also indexed in terms of the different ways in which scores can be generalized across six major universes: (1) scorer, (2) item, (3) time, (4) setting, (5) method, and (6) dimension (see Figure 7.3). The basis for this framework is generalizability theory as discussed by several authors (Cone, 1977, 1978; Cronbach, Gleser, Nanda, & Rajoratnam, 1972; Jones, Reid, & Patterson, 1975; Wiggins, 1973). Scorer generality refers to the extent to which data obtained by assessor (or scorer) are comparable to the mean of the observations of all scorers that might have been observing the behavior. Essentially, this concern is one of the agreement between assessors on observations of behavior. When two individuals agree, scores are said to generalize across the scorers.

Item generalization refers to the extent to which a given response is representative of those of a larger universe of similar responses. In behavioral assessment item generalization could be used in self-report instruments as when scores on odd-numbered items parallel those of even-numbered ones. Moreover, in behavioral observation odd-even scores might be compared during various phases of baseline and treatment assessment.

Generalization across time refers to the extent to which data collected at one point in time are representative of data that might have been collected at other times. Generally, behavioral assessors are concerned with the consistency of behavior across time, particularly within the context of stability in an intervention program.

Setting generality refers to the extent to which data obtained in one situation are representative of those obtainable in others. A behavioral assessor would be concerned with the degree of generality of a behavior across settings, such as from Classroom A to Classroom B.

The method generality of assessment refers to the extent to which data from different methods of measuring a target behavior produce consistency. Cone (1977) notes, "the method universe of generalization deals with the issue of the comparability of data produced from two or more ways of measuring the same behavioral content" (p. 420). Here behavioral assessors might be concerned with the general correspondence between measures of self-report and direct observation of behavior.

Dimensions generalization refers to the comparability of data on two or more different behaviors. When scores on a particular measure of one behavioral dimension relate to scores on other variables for the same clients, the scores are said to belong to a common universe.

Behavioral Assessment Methods

Our discussion will now focus on the most common methods of behavioral assessment, including (a) behavioral interviews, (b) self-report, (c) problem checklists and rating scales, (d) analogue measures, and (e) direct observation procedurs. In addition to this list, psychophysiological procedures, criterion-referenced testing, and more traditional psychoeducational testing are discussed within the context of their use in behavioral assessment.

While any one of these procedures might be used to assess a child's learning problems, this would be a rare instance. More likely, it is some combination of procedures and devices that provide an adequate data base for intervention. Moreover, it is the novel application of psychoeducational assessment procedures rather than their routine application that will provide useful information for educational programming. In this regard, it is the routine and stereotyped battery of assessment procedures that will likely lead to erroneous conclusions about intervention.

Interview Assessment. The interview method of gathering data is perhaps one of most common methods used in behavioral assessment. The interview assessment method has also been used widely in traditional psychotherapy and education (e.g., Benjamin, 1974; Bingham, Morre, & Gustad, 1959; Fear, 1973; Grant & Bray, 1969; Kahn & Cannell, 1957; Matarasso & Wiens, 1972; McCormick & Tiffin, 1974; Sullivan, 1954; Ulrich & Trumbo, 1965). Behavior assessors have also regarded the interview as an important clinical assessment technique (e.g., Ciminero, 1977; Ciminero & Drabman, 1977; Mash & Terdall, 1981; Meyer, Liddel, & Lyons, 1977; Linehan, 1977; Marholin & Bijou, 1978; Morganstern, 1976). However, even with interest in this area, concerns have been raised over the reliability and validity of the technique. Ciminero and Drabman (1978) noted "...the data available at this time suggest that we must be very cautious, if not skeptical, of interview data for children and parents" (p. 56).

This conclusion appears warranted, especially in light of the paucity of research on behavioral interviewing and the _informal strategies by which behavioral interviews are commonly conducted_ (Kratochwill, 1982). The lack of research on interviewing is generally well known (cf. Bergan, 1977; Ciminero & Drabman, 1978; Linehan, 1977). Also, while some systems present a conceptual framework for the behavioral interview (e.g., Kanfer & Grimm, 1977; Holland, 1970; Kanfer & Saslow, 1969), few formal script guidelines are provided and the assessor usually does not have a format for what specific questions should be asked at what point.

The compilation of data during the interview should yield a good basis for decisions about the areas in which intervention is needed, the particular targets for further assessment, some tentative targets for intervention, methods, and goals.

The interview can provide one of the first contact points for providing descriptions of learning and behavior problems, identification of specific behaviors needing modification, as well as variables (antecedent and consequences) controlling learning and social behaviors. A major contribution of what should be covered in a behavioral interview was presented by Kanfer and Saslow (1969). The authors noted that their guidelines can provide not only the initial information collected from the client, but also data relevant to formation of a treatment plan (see also Meyer, et al., 1977, for a similar proposal). [An outline of the approach presented by Kanfer and Saslow (1969, pp. 430-437) includes the general components that are useful for assessment of learning and behavior problems.]

More recently, Kanfer and Grimm (1977) proposed a differentiation of controlling variables and behavioral deficiencies into categories that can be matched with various intervention strategies. Their five categories include: (1) behavior deficiences, (2) behavioral excesses, (3) inappropriate environmental control, (4) inappropriate self-generated stimulus control, and (5) problematic reinforcement contingencies. The authors further indicate for each category: (a) briefly which kind of statements serve to define a particular behavioral problem as a member of each category, (b) examples of commonly encountered target behaviors in a class, and (c) briefly what therapeutic variables are available for change. Like many conceptual systems for interviewing, specific strategies for intervention can be found in the applied literature (e.g., Bandura, 1969; Goldfried &

Davidson, 1976; Kanfer & Phillips, 1970; Kanfer & Goldfried, 1975; Mahoney, 1974; Sulzer-Azaroff, & Mayer, 1977; Thoresen & Mahoney, 1974; Rimm & Masters, 1974).

Only one behavioral system has been developed for comprehensive interviewing of clients and consultees (care providers), namely, the Behavioral Consultation Model developed by Bergan and his associates (cf. Bergan, 1977). The Behavioral Consultation Model (cf. Bergan, 1978; Bergan & Tombari, 1975, 1976; Kratochwill & Bergan, 1978a, 1978b) provides a format to formalize the verbal interactions occurring during behavioral interviewing. The problem-solving model developed by Bergan and associates is designed to assist teachers and parents to define various problems (e.g., academic and emotional), to formulate and implement plans to solve problems (i.e., behavior intervention programs), and to evaluate various treatment goals (target of the interventions) and the effectiveness of educational programs.

The consultation interview format is actually a conceptual system for solving a variety of problems through an interview methodology. In this regard, the approach is particularly useful in psychoeducational assessment of learning and behavior problems. Consultive problem solving may focus on the achievement of long-range developmental goals, or it may center on specific concerns of immediate importance to the child-client and/or consultee. Developmental consultation focuses on behavior change that typically requires a relatively long period of time to attain. This form of consultation may require repeated interviews and the focus on subgoals which are subordinate to long-term objectives. Thus, repeated applications of the problem-solving process would be necessary until all the objectives of developmental consultation are achieved. This form of consultation is possibly more necessary in treating severe learning and behavior problems. For example, a child experiencing severe failure in reading (three or four years behind grade level) could possibly involve months of extensive intervention within this model.

On the other hand, many educational and psychological problems presented to the professional call for intervention on a limited number of specific behaviors of immediate concern to the teacher or parents. Bergan (1977) described consultation problems of this kind as problem-centered consultation. For example, consider a relatively specific problem of a child experiencing a high frequency of errors of orientation and sequence in handwriting. The majority of the child's words were written as mirror images of the correct word. During the plan implementation phase of the interview sequence, the teacher was requested to say "right" and praise after each correct response (i.e., writing a word correctly) and "wrong" and give corrective feedback after each incorrect response. After several repeated applications of this treatment, the child's writing reversed to normal patterns. Thus, the consultant's task was completed with the successful change in handwriting.

There are four stages in the consultation problem-solving model: Namely, problem identification, problem analysis, plan implementation, and problem evaluation. These stages (listed In Figure 7.4) describe the steps necessary to move from an initial designation of the problem through the plan development and implementation to achieve problem solution, to the evaluation of goal attainment and plan effectiveness.

Problem identification. In problem identification the problem or problems to be solved are specified. A problem is defined in the context of a discrepancy between observed behavior and desired behavior (Kaufman, 1971). For example, a child may know only three of the 26 letters of the alphabet. The problem is to devise a teaching strategy so that the remaining letters will be acquired.

1. Problem identification is achieved primarily by means of a problem-identification interview (PII). In the interview, the consultant assists the consultee to describe the problem of concern to him/her. In the case of a child who has not learned his/her letters, the consultant might say "tell me what Jack does when you present him with a letter to be learned." The question is deliberately phrased so that a socialization agent (e.g., teacher) will provide a rather specific description of the problem rather than a global one.

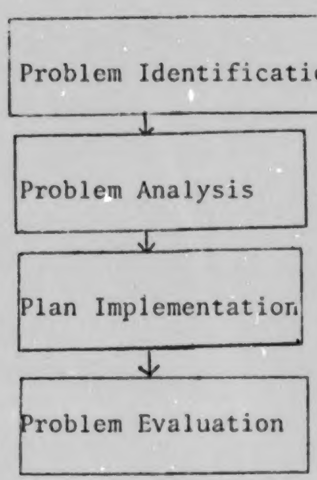
Problem Identification
Problem Analysis
Plan Implementation
Problem Evaluation

Figure 7.4 - States in consultative problem solving (Source: Bergan, J.R., Behavioral Consultation, Columbus, Ohio: Charles E. Merrill, 1977. Reproduced by permission).

The forms of the problem may shift throughout the PII, so that all relevant concerns are identified. Thereafter, a discussion of baseline measures which will provide a level from which to evaluate the treatment is outlined. In our letter learning illustration, a discussion of how to measure learning or its absence would be described (e.g., number of trials, responses, or presentations).

2. Problem Analysis. After a problem has been identified, consultation then focuses on problem analysis. The purposes of problem analysis are to identify variables that facilitate a problem solution and development of a plan to solve the problem specified in the problem-identification phase of consultation. Problem analysis is again primarily accomplished through a problem-analysis interview (PAI). In the PAI the consultant and consultee discuss client skills and environmental factors that might be controlling client behavior. For example, in our letter learning situation the consultant might suggest some behavioral principles to assist the teacher in teaching the letters. It might be determined that feedback and reinforcement need to be presented in a consistent fashion, or a discrimination procedure developed with similar letters. Subsequently, a specific plan would be developed to implement the suggested procedures. Such a plan might specify the conditions, time, place, and factors that facilitate generalization, and so forth.

3. Plan Implementation. The plan implementation phase of consultation is designed to implement and monitor the plan

designed during problem analysis. Data collection typically continues so that the consultee will have some indication as to the effectiveness of the plan. The only interviews that may occur during this phase are those used to check briefly with the consultee to determine that there is agreement between the plan specified and the plan implemented, and to deal with unforseen implementation problems. For example, it might be discovered that the parents are also working with the child in a fashion that serves only to have the two instructional strategies working in opposition to each other. The consultant must then deal with this problem.

4. Problem Evaluation. Problem evaluation takes place through a formal problem evaluation interview (PEI) and is conducted to determine if problem solution has been achieved by comparing data collected during plan implementation with the level of acceptable performance specified in problem identification. Moreover, consultation may be terminated if goals have been met (e.g., if the child acquires his letters). However, other problems may be introduced, and consultation may take on the developmental orientation. For example, it might be determined that the child has learning problems in math and other areas of reading. Consultation may then move back to problem analysis and the phase sequence continues.

There are several features that set aside the consultation interview system from other interview procedures described in the behavior therapy literature. First, the model conceptualizes the interview within the context of a consultant-consultee relationship. Thus, it is implied that indirect service will be provided to a client through some mediator (see also Tharp & Wetzel, 1969). Second, in contrast to the S-O-R-K-C sequence presented earlier (cf. Kanfer & Phillips, 1970; Kanfer & Saslow, 1969), the consultation interview system places less emphasis on the 0, or biological condition of the organism. However, it should be noted that such factors can be coded in the Bergan (1977) procedures. Third, the consultation interview system was primarily designed to be employed for academic and social problems, whereas the other strategies suggested in the literature have been primarily aimed at behavioral/social problems. Finally, and perhaps the distinctive feature of the behavioral consultation interview system is that specific and detailed coding systems have been developed for verbal interactions occuring during the actual interview (cf. Bergan & Tombari, 1975). Thus, the consultation-analysis technique enables the professional to assess the types of verbalizations emitted in consultation interviews. Since this is an important feature, it is briefly described here.

5. Message Clarification. The classification system developed by Bergan and his associates (e.g., Bergan & Tombari, 1975) is intended to articulate to the four-stage problem-solving model described above. The analysis system classifies verbal interchange in terms of four categories: source, content, process, and control. Table 7.5 shows

these four categories and the subcategories associated with them. The message source category indicates the person speaking. Content refers to what is being talked about. Proccess indicates the kind of verbal action conveyed in a message, and control refers to the potential influence of a verbalization by one participant in the interview on what will be said or done by another participant (see Bergan, 1977, pp. 30-46).

To code events of observation in accordance with the message classification categories, the behavior consultant employs a consultation-analysis record form (see Figure 7.6). The consultation-analysis record calls for coding in all four message-classification categories for each event of observation. The system is quite complex and requires extensive training (Brown, Kratochwill, & Bergan, 1981). It is a useful procedure for psychoeducational assessment and the most sophisticated interview procedure available to date. Although the model clearly represents a comprehensive assessment system within behavioral psychology, it also links assessment to treatment. Since consultation is largely a matter of verbal interchange between a consultant and

Table 7.5

TABLE 7.5 Message classification categories and subcategories

Categories	Message Source	Message Content	Message Process	Message Control
Subcategories	Consultant	Background Environment	Specification	Elicitor
	Consultee	Behavior Setting	Evaluation	Emitter
		Behavior	Inference	
		Individual Characteristics	Summarization	
		Observation	Validation	
		Plan		
		Other		

Table 7.6
Consultation Analysis Record Form

CONSULTANT ____________________ CASE NUMBER __________

CONSULTEE ____________________ INTERVIEW TYPE __________

PAGE __________

CONSULTATION-ANALYSIS RECORD

	Message Source		Message Content							Message Process							Message Control	
	Consultee	Consultant	Background Environment	Behavior Setting	Behavior	Individual Characteristics	Observation	Plan	Other	Negative Evaluation	Positive Evaluation	Inference	Specification	Summarization	Negative Validation	Positive Validation	Elicitor	Emitter
1																		
2																		
3																		
4																		
5																		
6																		
7																		
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		
25																		

(From J. R. Bergan & M. L. Tombari, The analysis of verbal interactions occurring during consultation. Journal of School Psychology, 1975, 13, 212. Reprinted by permission of Human Sciences Press, 72 Fifth Avenue, New York, New York 10011. Copyright © 1975.)

consultee and/or client, emphasis has been placed on the analysis of verbal behavior. Bergan (1977) suggests that consultant control of verbal behavior during consultation necessitates not only recognition of the types of verbal utterances that have occurred during interviews, but also the ability to produce different kinds of verbalizations to meet specific interviewing situations and problems. If a consultant is trying to elicit information about conditions controlling client behavior, he or she must be able to produce the type of verbal utterance most appropriate for the particular goal.

The future will likely see increased sophistication and use of the interview assessment stragety for purposes of nonbiased assessment (Reschly, 1981). Despite the recognized limitations of the interview procedure, several positive agents of interview assessment approaches have been identified (Linehan, 1977, pp. 33-34). First, interview assessment is a flexible means of obtaining data in that it can be used to gather both general information coverning many areas of the child's functioning and detailed information in specific problem areas. Second, variations in the careprovider's nonverbal and verbal behavior can be examined in relation to the assessor's questions thereby allowing an analysis of responding and lines of further inquiry. Third, the interview typically promotes the development of a personal relationship (in contrast to such methods as di observation where there may be no interaction between assessor and careprovider). Fourth, the interview may allow for potentially greater confidentiality relative to some other assessment procedures (e.g., rating methods, direct observation). Fifth, interview assessment may be an important source of gathering information from individuals who are unable to provide information through other means (e.g., those persons with limited communication skills, mental retardation). Sixth, the interview allows the assessor to modify his/her questions and responses to fit the person's conceptual/language system and affords an opportunity for modification of the interviewee's verbal description. This advantage must be balanced against the potential disadvantage of a nonstandardized script which may promote subjective interpretations.

There are, however, a number of issues that must be taken into account in the use of interview assessment strategies. Behavior assessors have long been skeptical of verbal reports, and with good reason. However, as Evans and Nelson (1977) have observed, by knowing some of the possible sources of error and bias in verbal reports it is possible to reduce distortions in material often unattainable by other means. They provided several guidelines when dealing with parental or adult informants (pp. 615-616).

Adult reports: Some cautions.

1. In written and verbal information, factual events in the child's developmental history are much more likely to be accurately reported than such

components as attitudes, feeling states, and child rearing practices.

2. Accuracy does not appear to be increased by repeated questioning, but can be improved by such devices as diagrams and by precise statements of the information required (e.g., McGraw & Malloy, 1941).
3. Poor recall is characteristic of information related to:
   (a) neonatal injuries or complications.
   (b) childhood illnesses (e.g., Mednick & Shaffer, 1963.
   (c) early attitudes regarding arrival of the baby (Brekstad, 1966), and
   (d) clinic referred behavior problems.
4. Length of time from the event to the interview does not influence the accuracy as much as the emotional significance of the event and the current level of anxiety (arousal) shown by the informant.
5. Distortions are likely to be in the direction of social desirability (e.g., placing the informant in a positive light, reporting socially accepted child-rearing practices).
6. Mothers may be more reliable informants than fathers, but when independent reports from mothers and fathers agree, the information is likely to be more valid.

7. General characteristics of accurate informants have not been identified.
8. There appears to be no information to suggest that social class or intellectual differences affect reliability of retrospective reports.

If the behavior assessor uses parental reports as evidence of the efficacy of a treatment program or other scientific conclusion, objective corroboration is required (cf. Allen & Goodman, 1966; Evans & Nelson, 1977).

Child Reports. Obtaining valid and reliable information from children also presents a challenge to the behavior assessor. Generally, we should not expect that children's descriptions of their own behavior to be any more reliable than their adult counterparts. In addition to the direct interviewing of major socialization agents of the child with learning and behavior problems, the child should be considered as an important source of information during interview strategies.

Although there are some technical guides for interviewing children (cf. Garrow, 1960), few behavior assessors have provided guidelines for this activity. Evans and Nelson (1977) suggest that there are three major types of information to obtain from an interview with a child:

1. information that only a child can provide regarding his/her perception of the problem;
2. likewise, information that only a child can provide regarding his/her perception of himself/herself; and,

3. indications of how well the child can handle himself/herself in a social situation with an adult.

An issue that typically arises from the behavior assessor is whether the child should be interviewed with his/her parents and/or teacher(s). On the one hand, separate interviews could lead to the child perceiving that adults are plotting some conspiracy, which may in turn make it harder to obtain accurate information during subsequent encounters. On the other hand, interviewing the parents and/or teacher and child together can provide valuable information. For example, the child provides the stimulus for certain questions and issues, as well as an opportunity to respond to certain points raised. Nevertheless, there will probably be occasion when a joint interview is aversive for parents, teachers, and child, because of the other's presence. Unfortunately, we have no empirical data to provide specific guidelines for such encounters. From the behavior assessor's perspective, it would be ideal to have an opportunity for all conditions (e.g., separate and joint interviews). Such opportunities would provide independent self-reported perception of the problem, establish some congruance among parties involved, conversely, establilsh some areas of noncongruance, and finally provide an opportuunity for each informant to have their "turn" at providing data relevant to the problem. However, from a time perspective, such options may not be possible, putting the burden on the assessor for the "best guess" as to which direction to go.

Interviewing the Child's Peers. It should also be mentioned that the child's peers, in addition to the child himself/herself, could be interviewed to gather relevant data. Peers can be especially reinforcing (or non-reinforcing), and may prove helpful in a functional analysis of the learning or behavior problem. Roff (1970) noted that an excellent predictor of adult maladjustment is a reputation as a child for being disliked by one's peers. Thus, children could be asked which children are having learning problems and what the possible causes are. There is also evidence to suggest that peer selection of children to fill a negative role in a hypothetical class play is a useful discrimination task (cf. Cowen, Pederson, Babigian, Rzzo, & Trost, 1973). Cowen et al. (1973) compared those adults appearing on a psychiatric register with matched controls on a large battery of tests and measures when these individuals were in the third grade. While the measures included standardized intelligence and achievement tests as well as personality measures and teacher ratings, the only measure that discriminated the psychiatric group from the controls was the negative role variable.

Self-Report and Behavior Checklists and Rating Scales. Self-report is an indirect assessment procedure because it represents a verbal description of more clinically relevant behavior occurring at another time and place.

Self-report assessment has sometimes been based as unreliable verbalizations in response to unstructured, open-ended questions. However, a variety of self-report inventories have been used to structure assessment (Bellack & Hersen, 1977). Behavior checklists and rating scales are conceptually similar indirect behavioral assessment strategies. In these strategies the child is asked to rate another person based upon past observations of that other's behavior. Due to the diversity of items that are included, the behavior of actual clinical interest (e.g., academic performance, social withdrawal) may or may not be involved. For example, a teacher may be asked to rate a series of behaviors in addition to the social withdrawal problem (e.g., fear, aggression, academic work). Presumably, other relevant educational problems may emerge from this assessment. Yet, the major feature of checklist and rating scale assessment strategies is that the rating occurs subsequent to the actual behavior of interest (Cone, 1977; Wiggins, 1973).

Self-Report. Due to the perceived problems inherent in subjective and unsystematic forms of self-report assessment, various inventories and schedules were developed (Tasto, 1977). Although self-report measures have generally been avoided by behavioral assessors, recent emphasis on cognitive processes (e.g., Kanfer & Goldstein, 1975; Thoresen & Mahoney, 1974) has focused attention on this form of measurement. Also, as Tasto (1977) has noted, in practice *"the operational criteria for the existence of problems are* self-reported verbalizations" (p. 154). For example, a child may report that he/she has an academic problem or has no friends. This report (a self-perception) is an important and relevant concern in assessment.

Self-report inventories are useful for at least two functions (Bellack & Hersen, 1977). To begin with, self-report measures can be useful in gathering data on motoric responses, physiological activity, and cognitions (see Figure 7.4). In any particular survey, the items may tap any of the three content areas or systems described above. Moreover, each of these questions, with the exception of cognitions (question 3) can be independently verified through the actual observation of behavior.

Another function of self-report measures is to gather data about a child's subjective experience. For example, one might ask a child "Do you like math?", "Do you dislike your peers?". It can be observed that this second set of questions include subjective components which are not objectively verifiable in the same way the first set is.

Numerous variables may influence the type of data one obtains from self-report and their correspondence to the actual criterion measure (usually the actual occurrence of behavior). Such factors as the source of the data will be important (e.g., written on verbal report by the client), the form of the questions asked, the content of the questions, situational factors, and operational specification of terms (Bellack & Hersen, 1977; Tasto, 1977; Haynes, 1978).

Behavior Checklists and Rating Scales. Many formal checklists and rating scales have been used in the educational and behavioral assessment of school age children. Walls, Werner, Bacon, and Zane (1977) provide a rather extensive catalogue of available scales, as have other authors (e.g., Severson, 1971). In many cases, behavioral assessors use scales originating from many different sources. As noted above, their use in behavioral assessment is premised on the nature of the data gathered and how such data are used in the development of an intervention program. Behavioral assessors using these procedures must, however, consider their indirect nature, avoid the hypothetical constructs sometimes associated with their use, and conduct a functional analysis in the natural environment once certain classes of behaviors are identified.

Several positive features of checklists and rating scales can be identified (Ciminero & Drabman, 1978; Kratochwill, 1982). First, checklists are typically economical in cost, effort, and assessor time. This is particularly the case in contrasting these procedures with direct observation of behavior in the natural environment. Second, many checklists are structured so that a relatively comprehensive picture of the problem can be obtained. However, such measures usually provide a very global picture of behavior. Third, due to the diverse range of questions asked in typical checklists and rating scales, the behavior assessor may be able to identify problems that were missed through other assessment methods such as in direct observation and interviewing. Fourth, data obtained from checklists and rating scales are usually (relatively) "easy" to quantify (as though factor analysis, multi-dimensional scaling, latent trait procedures). In this regard, they have been useful for classification of various behavior disorders (cf. Quay, 1979). Fifth, checklists and rating scales frequently may provide a useful measure for pre-and post-test evaluation of an intervention program. Sixth, such measures are frequently a convenient means of obtaining social validity data on therapeutic outcomes (cf. Kazdin, 1977; Wolf, 1978).

A number of considerations must be taken into account in the use of rating scales and checklists. Conceptual and methodological issues have been raised over their use in both research and practice (e.g., Anastasi, 1976; Ciminero & Drabman, 1978; Evans & Nelson, 1977; Kratochwill, 1982; Severson, 1971; Spivack & Swift, 1973; Walls et al., 1977).[7] A major problem with these procedures is that they represent an _indirect dimension of assessment_. Since data are gathered retrospectively, their relation to actual occurrences of the target behavior in the natural environment may be less than perfect. Second, while it appears that rating scale constructions have some criteria for generating items included in the scale, the rational may not be evident or remains unspecified. In this regard it is not always clear how items may relate to each other. Third, it is frequently unclear under what conditions the scale should be administered (e.g., at what time after observing the behavior). Fourth, there is often no clear rationale for the manner in which rating scale constructors rate the presence or absence of a particular kind of behavior and the kinds of categories employed to code various behavior scales and checklists are also characterized by considerable variation within a particular scale with regard to the kinds of judgments required. Sixth, a large number of rating scales are constructed to detect the presence of negative behaviors or problems (i.e., behavioral excesses and deficits) and less frequently focus on positive behaviors (assets). Finally, many published scales fail to meet standards for reliability, validity, and norming (cf. Walls et al., 1977).

Rating scales and checklists will likely continue to be used extensively in behavioral assessment in schools and other applied settings. A continuing reason for their popularity typically relates to the general ease with which such devices are administered (but not necessarily interpreted). Nevertheless, the rating scale and checklist user should consider the aforementioned conceptual and methodological limitations if he/she wishes to reduce bias in the assessment process.

Self-Monitoring. Self-monitoring refers to the act of a child in which some occurrence(s) of his/her behavior are discriminated and then recorded. This procedure is regarded as a direct assessment procedure in that behavior is recorded at the time of its actual occurrence. Several major sources have provided a review of the applications of self-monitoring (e.g., Ciminero, Nelson, & Lipinski, 1977; Haynes, 1978-Chapter 9; Kanfer & Phillips, 1970; Kazdin, 1974; Mahoney, 1977; McFall, 1977; Nelson, 1977; Watson & Tharp, 1974; Workman & Hector, 1978). Self-monitoring (SM) can be used for both assessment and treatment of various problem behaviors. Its use in assessment and treatment involve somewhat different considerations.

Self-Monitoring Assessment. When SM assessment is employed, data on the child's behavior are useful for at least two reasons. First, the client may be requested to SM during the initial stages of educational assessment when the professional is attempting to identify specific problems. In this regard, baseline response levels help verify the existence of a problem. SM may also be used to gather information on how successful the intervention program is. The range of application of SM to various target behaviors has been quite extensive and the interested reader is referred to the references listed above for examples.

Many different recording devices and methods have been used for SM assessment. Some of the more common include record booklets, checklists, forms, counters, timers, meters, measures, scales, residual records (e.g., empty pop bottles), archival records (e.g., telephone bills), diaries, and many others.

When SM is used for assessment, a number of variables influence the reliability and validity of the data. Both the accuracy and reactivity of SM have been identified as factors influencing the data (see Table 7.7). The accuracy of SM depends on the following 10 factors (McFall, 1977, pp. 200-201):

1. Training. Children should be trained in the use of SM. Training will generally result in better accuracy and increase the credibility of assessment.
2. Systematic methods. Systematic SM methods will usually result in more reliable and accurate measures than those that are more informal and nonsystematic.
3. Characteristics of the SM device. A SM device which allows simple data collection, and which does not depend heavily on the child's memory will usually provide more accurate data in assessment.
4. Timing. In general, the closer in time the actual SM act is to the occurrence of the target behavior, the more likely the data will be accurate.
5. Response Competition. When a child is required to monitor concurrent responses, his/her attention is divided. This may cause interference and thereby reduce the accuracy of the SM data.

Table 7.7

Factors Influencing Self-Monitoring Assessment

	Factor	
	Accuracy	Reactivity
Dimensions	1. Training	1. Motivation
	2. Systematic methods	2. Valence
	3. Characteristics of monitoring device	3. Target behaviors
	4. Timing	4. Goals, reinforcement, and feedback
	5. Response competition	5. Timing
	6. Response effort	6. Self-monitoring devices
	7. Reinforcement	7. Number of target gehaviors
	8. Awareness of accuracy assessment	8. Schedule of self-monitoring
	9. Selection of target	
	10. Characteristics of clients	

6. Response effort. The more effort (i.e., time and energy) the child must spend on the SM activity, the less accurate the data may be.

7. Reinforcement. Contingent positive reinforcement for accurate recording will usually increase accuracy. Some external criterion can also be established for accuracy improvement.

8. Awareness of accuracy assessment. The professional should monitor the child's data and make him/her aware that accuracy is being monitored. Such awareness will usually increase accuracy.

9. Selection of target behaviors. Since some behaviors are more salient, more easily discriminated, or more memorable, variations in accuracy will occur as a function of these dimensions. Generally, higher levels of accuracy have been established on motor behaviors (e.g., head touches) than verbal behaviors (e.g., number of times the person says "you know") and positively valued behaviors are more accurately recorded than those that are negatively valued.

10. Characteristics of child. Some children are more accurate recorders than others. One would generally expect young children to be less accurate than older children, adolescents and adults. However, individual variations will occur within ages.

Reactivity may also be problematic when SM is undertaken because unintended or unwanted influences caused by the act of recording yield data that are not representative had SM not been used. McFall (1977, pp. 202-204) presented eight variables that should be considered:

1. Motivation. Children who are motivated to change their behaivor prior to engaging in SM are more likely to demonstrate reactive effects.
2. Valence. Depending on how children value a particular SM behavior, it may or may not change. Generally, positively valued behaviors are likely to increase, negatively valued behaviors are likely to decrease; and neutral behaviors may not change.
3. Target behaviors. The nature of the target behavior for SM may influence reactivity. Also, the number of target behaviors monitored at one time may produce different reactive effects. Sometimes, two behaviors being monitored may be more reactive than one.
4. Goals, reinforcement, and feedback. Specific performance goals, feedback and reinforcement scheduled as part of SM will increase reactivity.
5. Timing. Reactivity may vary as a function of the timing of SM. As the time between the natural occurrence of a behavior and the recording of the behavior increases, reactvitity may decrease.
6. Self-monitoring devices. Generally, the more

obtrusive the recording device, the more reactive it tends to be (e.g., a hand held timer is more reactive than one that is out of sight and "awareness").

7. Number of target behaviors. As the number of target behaviors being monitored increases, reactivity may decrease.

8. Schedule of self-monitoring. Continuous SM may be more reactive then intermittent SM.

The 10 accuracy variables and the eight reactivity variables may be problematic in assessment. When SM is used as an intervention somewhat different concerns must be considered.

Self-Monitoring as an Intervention. Self-monitoring is frequently used as a therapeutic technique and it often has been used as one component of a more complete system of behavioral self-control (Thoresen & Mahoney, 1974). A total program might include the following components: (a) self-assessment where the child examines his/her own behavior and determines whether or not he/she has performed certain behaviors; (b) self-monitoring, (c) self-determination of reinforcement wherein the child determines the nature and amount of reinforcement he/she should receive contingent upon the performance of a given class of behaviors, and (d) self-administration of reinforcement wherein the child dispenses his/her own reinforcement (self-determined or not) contingent upon performance of a given class of behaviors

(Glynn, Thomas, & Shee, 1973).

When SM is used as an intervention, accuracy and reactivity take on quite different roles. Accuracy plays a minor role in fostering therapeutic change since regardless of whether or not children monitor accurately, SM may produce a positive behavior change. While reactivity is something to minimize in assessment, it is usually fostered to maximize therapeutic change. Not all reactivity may be therapeutically desirable and the professional must arrange conditions so as to facilitate positive reactive change.

Positive Characteristics of SM. Despite some potential methodological limitations, SM may be advantageous for behavioral assessment for several reasons. First, it is a relatively cost-efficient means of assessment relative to such techniques as observational assessment. However, the professional must take into considerations such factors as the training time and monitoring in such a cost analysis. Second, SM may be the only assessment option, as in measurement of private behaviors (thoughts). Third, SM can minimize the sometimes obtrusive effects of assessment that occur with other assessment procedures (e.g., interview, direct observation). Fourth, SM can help verify the existence of a problem in combination with other assessment methods.

Analogue Assessment. Another direct assessment procedure requires clients to respond to stimuli that simulate or approximate those found in the natural environment. In such assessment analogues the child is usually requested to role play or perform as if he/she were in the natural environment. Analogue assessment procedures have been used for many years within behavior therapy, but it is only recently that systematic features have been outlined and advantages and disadvantages considered (cf. Haynes, 1978; McFall, 1977; Nay, 1977). Relative to direct naturalistic assessment, analogue methods offer several positive contributions. First, particularly in research, they permit increased opportunities for experimental control. This positive feature may also emerge when analogue assessment is being used for clinical purposes. Many variables operating in the natural environment contaminate assessment efforts and a more analogue assessment may reduce these. In this regard the professional may be able to gain a good perspective on the problem free from some of the contaminating factors usually present in the natural setting (e.g., classroom). Second, analogues may reduce the amount of distortion that sometimes occurs when an observer is present in naturalistic settings. Third, analogue assessment may allow assessment of behaviors which are impossible to monitor in naturalistic settings. Fourth, relative to direct observational assessment procedures, analogue strategies may be less costly on several dimensions. Fifth, analogue assessment may help simplify and reduce complex problems. Through analogue assessment we may be able to control extraneous influences, isolate and manipulate specific variables, and reliably measure their effects. Sixth, analogue assessment procedures may help professionals avoid certain ethical problems that emerge in naturalistic observation. Thus, under analogue assessment conditions the professional may be able to test a procedure to learn about its characteristics prior to implementing it in the natural environment.

Five categories of analogue methods have been identified by Nay (1977); paper and pencil, audiotape, videotape, enactment, and role play analogues. Paper and pencil analogues require the child to note how he/she would respond to a stimulus situation presented in written form. For example, teachers may be asked to respond to a series of multiple choice questions which depict different options to follow in implementing behavior management procedures. In paper and pencil analogues the stimulus situations are presented in a written mode with responses options written, verbal, and/or physical. The child is usually presented the stimulus and a cue for a response is made. The response made may be verbal in that the child is asked to describe what he/she would do and/or physically respond as he/she typically would. While a major advantage of these procedures is that they can be given to large numbers of children at the same time and that they are easily quantified, the predictive utility of these procedures usually remains unknown. Moreover, this type of measure is limited because the professional does not observe overt behavior in response to the actual stimulus.

Audiotape analogues present stimulus items in some type of auditory format. Some characteristics of these procedures include a set of instructions to the child and a series of audio situations presented by the professional. The child is typically required to make a verbal or other physical response. For example, the professional may present audio transcripts of a teacher presenting information to a class of school age children. The child may be requested to respond through role play or free behavior. Although the audio analogue shares many of the advantages of the paper and pencil analogue, it may not approach realistic stimulus conditions.

The videotape analogue uses video technology to present a relatively realistic scene for the child. In this regard it can closely approximate the naturalistic setting. Most often both audio and visual components are used. Video analogues can also be used for training-intervention, as in the teaching of social or academic skills. Cost and availability of the video equipment represent major limitations of this procedure.

Enactment analogues require the child to interact with relevant stimulus persons (or objects) typically present in the natural environment within the contrived situation. Sometimes the professional may bring relevant stimulus persons (e.g., peers, teachers) into the assessment setting to observe child responses, as has been done in assessment and treatment of selective mutism (Kratochwill, Brody, & Piersel, 1979). A major advantage of this approach is that stimuli can be arranged to be nearly identical to the natural environment. Yet, a limitation of this procedure is that the situation may still not duplicate the natural environment.

The role-play analogue can be used within the context of any of the aforementioned assessment procedures. Sometimes a script is presented and the child is asked to covertly rehearse or overtly enact certain behaviors under various stimulus situations. A professional may ask a student to role-play asking a teacher's assistance to assess various preacademic skills. The child may play himself/herself or someone else. Specific instructions may be present or absent. Flexibility in format is a major advantage of this procedure as is the option for direct measurement of the behavioral responses. As is characteristic of other analogue assessment procedures, a major disadvantage is the potential lack of a close match between the analogue and the natural environment.

The analogue assessment procedure presents many behavior assessment options in educational settings. Nevertheless, both reliability and validity issues need to be addressed when these procedures are used (Nay, 1977). Therapists employing these procedures should assess reliability data on target responses. A check on the validity of the analogue is made by comparing the contrived assessment with the target behaviors occurring in the natural environment. As is true of other assessment procedures, analogue assessment may best be used as one of several techniques to assess behavior.

Direct Observational Assessment. Direct observational assessment is a most commonly used procedure in behavioral research and practice. Jones, Reid, and Patterson (1975) summarized three major characteristics of a "naturalistic observational system", including recording of behavioral events in their natural settings at the time they occur, not retrospectively; the use of trained impartial observer-coders, and descriptions of behaviors which require little if any inference by observers to code the events" (p. 46).

Observational assessment strategies are commonly affiliated with behavioral approaches (e.g., Johnson & Bolstand, 1973; Jones et al., 1974; Kent & Foster, 1977; Lipinski & Nelson, 1974) but are not limited to this orientation. They are used in rather diverse areas of psychology and education (e.g., Boehm & Weinberg, 1977; Cartwright & Cartwright, 1974; Flanders, 1966, 1970; Hunter, 1977; Lynch, 1977; Medly & Mitzel, 1963; Rosenshine & Furst, 1973; Sackett, 1978a, 1978b; Sitko, Fink, & Gillespie, 1977; Weick, 1968; Weinberg & Wood, 1975; Wright, 1960).

The rather extensive literature in this area does not allow a thorough presentation (see Haynes, 1978 for more detailed coverage in behavior therapy). When used in educational and psychological assessment, many issues emerge over the utility of these procedures . One major issue in its use in clinical assessment is the distinction between observational procedures and actual observation instruments (cf. Kratochwill, et al., 1980). Most professionals have used some type of observational procedures in their assessment work. This may take the form of their direct observation of a child in a classroom or having a parent or teacher record the occurrence of some behavior. Figure 7.5 presents an example of a record form used by a special education teacher who had an aide observe a child over a one week period. While observational measurement procedures may vary considerably under a number of dimensions (e.g., the person observing, the target response, the sophistication of the form), they are most commonly used as part of a more general assessment battery.

In contrast to these observational procedures, there are relatively few specific observational instruments in use in behavioral assessment. The paucity of instruments for direct observational assessment is likely due to the lack of attention to the development of these scales and the need to design assessment forms for specific situations and problems (Mash & Terdal, 1981).

Among the instruments that have been developed, most focus on a rather specific range of behaviors (e.g., arget Client-----(name)---------------------------------- Date:----------

etting:----------(place)------------------------------

Observation Period (Seconds)

inutes:			1				
econds:	10	20	30	40	50	60	10
arget ehavior	T	T	TO	To TO	To TO	$T_1{}^0$	
ositive Behaviors							
eer Attention							

Target Behavior:----------(definition)----------------------------------

ositive Behaviors:-------(definition)----------------------------------

Peer Attention:-----------(definition)----------------------------------

Figure 7.5 - An example of a multiple behavior recording format used for direct observational assessment. Numbers across the top of this sample block indicate 10-second observe, 10-second record intervals. Target behaviors for the child, and peer are listed down the left margin Problem behavior (target) for the child are coded as Talking (T), throwing objects (TO) and out of seat (O). Problem and desirable child behaviors are mutually exclusive in any one 10-second interval.

Aleviros, DeRissi, Liberman, Eckman, & Callahan, 1978; O'Leary, Romanczyk, Kass, Dietz, & San Tagrossi, 1971; Patterson, Ray, Shaw, & Cobb, 1969; Wahler, House, & Stambough, 1976). Each of these coding systems is used in different settings, including institutional program evaluation (Aleviros et al., 1978)[8], home (Patterson et al., 1969)[9], school (O'Leary et al., 1971)[10] and home and school (Wahler et al., 1976)[11]. Each of these systems represents a promising observational instrument for assessment in research and practice (see Ciminero & Drabman, 1977 for a brief overview).

Direct observational measurement is usually the preferred method of assessment in therapy and research. Yet, the number of methodological issues that have been raised in recent years has made this assessment procedure complex[12] (cf. Ciminero & Drabman, 1977; Gelfand & Hartmann, 1975; Haynes, 1978; Johnson & Bolstad, 1973; Kazdin, 1977; Kent & Foster, 1977; Wildman & Erickson, 1978). From the available literature has come some recommendations that can make this form of assessment more credible in practice (some specific factors that bias this form of assessment were reviewed in Chapter 5). First, individuals functioning as observers should be well-trained. Training should include samples of behavioral sequences and environmental settings which closely resemble the behaviors and settings in which actual data collection will occur.

Second, two or more observers should be involved in assessment efforts to establish interobserver agreement on the response measures. Observers should be trained together and the scores compared with a single formal criterion, and training should be long enough to ensure that there is agreement to a specified criterion on each code.

Third, the conditions for assessing observer agreement should be maintained to insure consistent levels of agreement. Continuous overt monitoring and covert monitoring may help generate stable levels of agreement (cf. Wildman & Erickson, 1978).

Fourth, observer bias can be reduced by not communicating the specific intervention plan to the observer(s). Possibly, explicit instructions to the observer indicating that the specific outcomes are unknown may be preferable to completely avoiding the topic.

Fifth, in the absence of instruments or coding sheets for a target problem, specific observational codes should be constructed so that behaviors can be easily rated. The professional should typically be conservative in the number of codes that are to be rated at any one time.

Sixth, operational definitions should be constructed for each specific behavior to be observed. Definitions should also be tested to ensure that two independent observers can obtain and maintain high levels of interobserver agreement.

Seventh, observations should be conducted in an unobtrusive fashion. To assist in the examination of obtrusiveness, data should be monitored for evidence of reactivity or bias.

Eighth, measurement of the generality of observational data across different settings should be conducted. While direct observations should take place in the settings in which the target behavior has been identified, multiple assessment across behaviors and settings will further elucidate the extent of the problem and help monitor intervention effects.

Finally, normative data are quite desirable in many cases and should be considered in observational assessment. Normative data can help objectively identify behavioral excesses and deficits in a given client (Hartmann et al., 1979; Nelson & Bowles, 1975).

While direct observational measures will likely remain an important procedure within behavioral assessment, much work remains to be done to make this form of assessment less expensive, less time consuming, and more versatile. Because this strategy involves less inference about a particular behavior relative to many traditional assessment practices, and because it emphasizes a repeated assessment of the child across various phases of intervention, it should be used as often as possible.

Psychophysiological Assessment. Behavior assessors have increasingly focused on psychophysiological measures of behavior in part due to the growing emphasis on the three response systems. Psychophysiological measurement is defined as "the quantification of biological events as they relate to psychological variables" (Kallman & Feuerstein, 1977, p. 329). The rise in interest in psychophysiological measures in child assessment is also due to increased sophistication in instrumentation, increased use of biofeedback and other behavioral procedures to treat phychophysiological disorders, and the finding that independent measures of physiological responding do not correlate perfectly with verbal reports and overt behavior, thus making it increasingly desirable that they be used (Ciminero & Drabman, 1978). It is generally recognized that psychophysiological assessment is in its very early stages of development relative to other assessment procedures. However, their use in schools has been particularly limited. Psychophysiological assessment in schools and other settings necessitates the consideration of several issues (Kallman & Feuerstein, 1977). First, due to their complexity and expense, physiological recordings should provide data that cannot be obtained as reliably and efficiently as by other procedures. Second, within educational settings, psychophysiological assessment should provide the professional with information about the selection and evaluation of an intervention strategy. Third, psychophysiological assessment procedures should possess adequate reliability and validity for their use.

Several classes of problems have been identified when physiological measures are used (Hersen & Barlow, 1976). These factors may reduce the usefulness of this form of assessment. First, equipment used to monitor physiological responses is sometimes characterized by mechanical failure. Second, the professional using physiological measurements must allow time for adaptation during various phases of assessment. Sometimes, physiological measurement is initially reactive and this effect must be eliminated so that the effects of the intervention can be separated from the effects of reactivity alone. When physiological responses are repeatedly measured, habituation and adaptation may be problematic. In this regard, the effect of the intervention must be distinguished from mere habituation or adaptation to recording. Third, various assessor and contextual variables may interact with the physiological measures. Fourth, when various physiological response systems (e.g., GSR, heart rate, blood pressure, etc.) are used as indices of emotional arousal, the specific emotion experienced by the child cannot be assumed to occur in the absence of a self-report confirmation from the child. Finally, there appears to be some evidence for individual differences in autonomic reactivity. For example, different peripheral autonomic systems may show low or inconsistent correlations across clients (Zuckerman, 1970).

As with other behavioral procedures, psychophysiological assessment can provide important information for the design of intervention programs. Yet, the increased use in school settings will likely be slow based on cost and efficiency considerations.

Criterion-Referenced Assessment

Criterion-referenced assessment has been closely aligned with but not limited to the behavioral paradigm (Bijou, 1976; Cancelli & Kratochwill, 1981). Since criterion-referenced tests were first introduced (Glaser & Klaus, 1962) continued clarification of the term as well as issues that must be addressed in its use have proliferated (cf. Hambleton, Swaminathan, Algina, & Coulson, 1978). In the early literature criterion-referenced tests were considered precise measures of highly specific discrete behavior capabilities. Such behaviors were purported to be hierarchically sequenced, as derived through task analysis procedures (cf. Gagne, 1961, 1968; Resnick, Wang, & Kaplan, 1973). Glaser (1971) provided an early definition of a criterion-referenced test:

> A criterion-referenced test is one that is deliberatly constructed to yield measurements that are directly interpretable in terms of specified performance standards. Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. Measurements are taken as representative samples of tasks drawn from the domain and such measurements are referenced directly to this domain for each individual (p. 41).

Within this conceptualization, the term domain-referenced test has evolved. Thus, whether one prefers the term criterion-referenced (Hambleton et al., 1978) or domain-referenced (Subkoviak & Baker, 1977), it is generally assumed that the concept of "domain" is implied. Nevertheless, these notions about criterion-referenced tests have evolved outside a behavioral orientation. It appears most useful to consider that performance on a criterion-referenced test is a function of the immediate test situation and the previous interactions that comprise the history of the child (Bijou, 1976). Specific responses to items on a criterion-referenced test may be due to (a) the nature of the test items, and (b) the setting factors in taking the test.

Behavioral assessors have noted that criterion-/domain-referenced tests are greatly improved with an empirical validation of homogenous item domains (e.g., Bergan, 1978; Cancelli, 1978; Dayton & Macready, 1976; Macready & Merwin, 1973). However until recently, procedures for establishing homogeneous item domains have not been available (e.g., latent structure analysis). With the development of procedures for empirically validating the scope and sequence of domains of homogeneous items, a new form of criterion-/domain-referenced assessment, labeled path-referenced assessment (Bergan, 1978, 1980) has been developed. This assessment procedure provides information about the client/learner which allows specific identification of skill and/or domain deficiencies as well as the sequence (i.e., "path") of curriculum instruction that will lead most efficaciously to mastery of the task identified .

Criterion-/domain-referenced tests have generally been used for three purposes within educational settings (Bijou, 1976): (a) to diagnose problem behavior, (b) to monitor learning, and (c) to assess readiness for placement in a prescribed educational program. A central theme in their use is that they measure a child's competence in a particular area and assist in the design of a specific instructional program. Yet, several criticisms of criterion-/domain-referenced assessment have emerged related to the lack of normative comparisons in the assessment activity (e.g., Ebel, 1970; Hofmeister, 1975). Based upon the common use of criterion-/domain-referenced assessment, such measures do not provide normative data - a characteristic deemed desirable by some professionals.

A response to this issue of normative data must take into consideration that norm-referenced and criterion-/domain-referenced tests are really designed for different purposes. Items on a criterion-referenced tests are typically selected randomly from each domain during test construction while psychometric theory governing the selection of items for norm-referenced devices suggests that in order to discriminate between good and poor learner's, items which are passed by half a sample of the population are best (Subkoviak & Baker, 1977). Individuals desiring normative information from criterion-/domain-referenced assessment should consider the use of social validation as an alternative to psychometrically established norms (cf.

Kazdin, 1977; Wolf, 1978). This extension involves social validity (Kazdin, 1977), a procedure which refers to assessing the social acceptability of some intervention. Wolf and his associates suggested that interventions be socially validated (e.g., Maloney, Harper, Braukmann, Fixsen, Phillips, & Wolf, 1976; Minkin, Braukman, Minkin, Timbers, Fixsen, Phillips, & Wolf, 1976; Phillips, Phillips, Wolf, & Fixsen, 1973; Wolf, 1976). Kazdin (1977) reviewed several facets of social acceptability:

> Initially, the acceptability of the focus of the intervention can be assessed. This aspect of social acceptability refers to whether the behaviors selected are important to individuals in the natural environment. Second, the acceptability of the procedures can be assessed. Presumably, many procedures might alter behavior (e.g., reinforcement of a particular response, time out, shock). Acceptability of, or consumer satisfaction with, the procedure can be determined and used as a basis for selecting among effective techniques. Finally, the importance of the behavior change achieved with treatment can be validated by examining the change in light of the performance of the nondeviant peers in the environment or through evaluations by individuals in everyday contact with the client (p. 430).

An important component of social validation involves determining whether behavior change is clinically relevant for the child. One way that this might be accomplished is in assessing the child's functioning in the environment after the academic performance has been achieved. In this case, validation of intervention effects can be accomplished in two ways: Namely, social comparison and subjective evaluation. Both of these involve somewhat different considerations and methods.

Social comparison requires that the professional identify individuals similar to the child in subject and demographic variables, but who differ in performance on the target behavior (e.g., knows the multiplication tables 1 through 123). Kazdin (1977) suggested two ways for this assessment to be conducted. First, assessment of a target behavior is determined to be deficient and therefore warrants an intervention. Second, the level of performance of peers who do not warrant an intervention could serve as the criterion for the intervention on the deviant child. Thus, if the intervention is effective, the child's academic behavior should fall within the normative level of peers. Research reviewed by Kazdin (1977) suggests that many applied intervention programs whose target behaviors involve social behavior patterns have successfully used social comparison studies. For example, O'Connor (1972) developed social interaction in nursery school children with modeling or modeling combined with shaping. Prior to treatment, isolate children were below the level of their non-isolate peers in such behaviors as proximity to others and visual and verbal contact with peers. After treatment, social interaction in the classroom of the trained children surpassed the level of their non-isolate peers, an effect that was maintained up to six weeks of follow-up. These results suggest that the magnitude of change was clinically important in that setting. There appears to be fewer studies to provide good examples of social validation of academic skills.

Subjective evaluation of intervention effects consists of judgments about the qualitative aspects of performance. Presumably the academic performance that has been altered can be observed by individuals who are in the natural environment with the child (teacher) or who are in a special position through training and professional skills (e.g., special education teacher, psychologist) to judge the behavior. This form of evaluation is quite common in applied behavioral research. For example, subjective evaluations have been used with reinforcement techniques designed to alter compositional responses of elementary school children (e.g., amount of writing, use of words not previously used, varied sentence beginnings). Subjective evaluations by adults including teachers or college students have revealed that compositions completed after training are rated qualitatively better than those completed during baseline (e.g., Brigham, Graubard, & Stand, 1972; Maloney & Hopkins, 1973; Van Houten, Morrison, Jarvis, & McDonald, 1974). Both social comparison and subjective evaluations can be employed to evaluate treatment effects.

Social validation represents an important alternative for those individuals wishing to evaluate the effects interventions in applied settings. Moreover, such evaluations provide an alternative to conventional norm-referenced tests. However, these procedures are not without their problems. Normative standards may be an inappropriate criterion against which to evaluate change. As Kazdin (1977) has noted, a goal might even be to change the normative level. For example, one of the authors (Kratochwill) has worked with teachers who argue that reading and many readiness skills should not be taught in kindergarten. In this situation the goal would be to achieve a new level that would be desirable for both the children and for teachers in later grades.

As Kazdin (1977) has noted:

> ...classroom applications might bring the academic performance of a student up to the level of his or her peers. While this would be a successful intervention in some sense, whether normative levels should ever serve as a standard might be questioned. Normative levels of academic performance in most classrooms can be readily accelerated with reprogramming teacher behavior and curricula (p. 439).

The same issue can, of course, be raised with any normative standard. The issue is that normative levels of performance as a criterion for evaluating change implies a satisfaction with these levels. Nevertheless, the issue is that many people working with the client in the natural environment would accept average behavior, especially if it has previously been deviant.

Identifying the normative group may also be difficult for some individuals. While it might be expected that a child of normal measured intelligence would be able to count from 1-100, this goal may be unrealistic for a mentally retarded child. It could also be somewhat arbitrary to specify those individuals who constitute the normative group. Should a Mexican-American child have the same normative group as his/her anglo peer? As Kazdin (1977) has noted, simply defining ones' "peers" or the normative group hinders many variables that might be relevant for judging intervention effects. The professional might want to take into account such factors as age, SES, IQ, and family environment.

Use of Traditional Assessment Devices in Behavioral Assessment. Much has been written on the limitations of traditional assessment practices, both within personality (e.g., Hersen & Barlow, 1976; Mischel, 1968) and ability testing approaches (e.g., Bersoff, 1973; Bijou & Grimm, 1975; Laosa, 1978; Kratochwill, 1977; Mann, 1971; Salvia & Ysseldyke, 1978; Ysseldyke, 1973). Behavior assessors have typically rejected various standardized tests of ability and have instead tended to argue for the use of criterion-referenced tests (cf. Livingston, 1977) and task analysis procedures (Mercer & Ysseldyke, 1978).

As was mentioned earlier, one of the aims of traditional tests is to predict which children might be in need of special education services. A major limitation of IQ tests (and other instruments used to diagnose learning problems) for the professional interested in developing intervention programs is that the constructors of these scales were really concerned with large group prediction. One of the main purposes of psychoeducational behavioral assessment should be with confounding these predictions by altering the child's problems. Binet developed the mental test as a screening device so that "feeble-minded" children could receive special education in the Parisian schools. However, what was significant is that no one seems to know how successful they were (i.e., did those children end up better off than just setting it out at the bottom of the regular classes?) Also noteworthy is that the research comparing the academic achievement of children in special education classes versus regular class has yielded equivocal results (e.g., Blatt, 1958; Cassidy & Standon, 1959; Goldstein, Moss, & Jordov, 1965). Moreover, it is the conceptual shift from prediction to potential that has obscured the disadvantaged, race, intelligence, compensatory education debate (Cronback, 1975b). Nevertheless, the notion of using the IQ score as a measure of "intellectual" potential is strong despite the fact that the predictive validity of the IQ score for most special populations is largely unknown.

What use can the behavior assessor have for the IQ test? Evans and Nelson (1977) suggest four qualities that might be considered: Namely, the standardization feautre, goal setting for remediation by behavioral methods, assessment of items not learned at school, and context of assesement. Others have noted some positive features of traditional IQ tests as well (e.g., Ciminero & Drabman, 1978).

Standardized Qualities. Evans and Nelson (1977) suggest that the standardization feature of tests allows definition of one's target population relative to others. This may be relevant in evaluating the outcome of intervention programs. A second point is that standardized test data also allow evaluation of the substantive significance of a behavioral program (cf. Nelson, 1974). This appears to be an improvement on some behavioral interventions that have tended to report outcome data in the form of changes on some arbitrary scale, the meaningfulness of which is unknown. Staats (1971; 1973) suggested that standardized test scores provide an additional source of data against which to evaluate statistically the success of a behavioral program. While the standardization sample could be conceptualized as a large control group, Evans and Nelson (1977) note that the statistical problems inherent in this strategy are considerable and would require knowledge of the reliability of the test for the population from which the treated children were drawn. Moreover, the test-retest reliabilities typically are low for the kinds of special children treated by behavioral procedures.

Goal Setting. A second advantage of standardized tests that has been raised is that, given well-constructed age norms, they can reveal an area of deficit and thereby help set academic goals for remediation by behavioral methods (Bijou, 1971). Evans and Nelson (1977) suggest two problems with this approach. First, one problem is to ascertain by how much a score on a particular subtest has to deviate before the child can be thought to have a serious deficiency in the area. The answer to this is again related to a statistical issue regarding error of measurement of the individual subtests and the scatter of the scores obtained. A second problem relates to the test item. A child who does poorly on visual sequential memory from the ITPA will likely be referred for training in "visual memory", but as we have noted above, the implication for reading instruction will likely be tenuous. Moreover, item content for prediction may not relate to instructional goals. Unfortunately, despite the fact that test constructors insist that IQ tests should not be used as tests of cognitive abilities (cf. Wechsler, 1975), they continue to be used for these purposes.

Item Content. Many standardized tests of ability (e.g., IQ) include/assess knowledge of items not learned exclusively at school. Thus, another use of traditional ability tests would be that they allow one to compare what the child has learned generally (as measured in an IQ score) with what he/she has learned at school (e.g., some score on a school achievement test) (Evans & Nelson, 1977). Evans and Nelson (1977) note that a statistically significant difference between the two measures and with the achievement measure lower, would suggest remediation of rather general classroom learning and studying skills. Unfortunately, there are some problems with this reasonable suggestion. First, there is some overlap between the two measures, so for a test of ability/achievement differences one would have to sort out specific items for further analysis. Second, a statistically significant difference may not be a meaningful difference. Finally, a point recognized by Evans and Nelson (1977) is that typical achievement measures are sometimes so general that remedial efforts could be quite misdirected even if they did correspond to a curriculum in the classroom, which they typically would not.

The argument that one could compare scores on one IQ test with those obtained from a more "culture reduced" test to estimate the degree of deficiency in skills specific to the dominant culture seems quite reasonable (cf. Evans & Nelson, 1977). For example, a test administered in both standard and nonstandard English could yield a discrepancy that would allow the professional to determine if the minority group child had a "cognitive deficiency" or a limited knowledge of standard English (Quay, 1971). Nevertheless, to hypothesize a cognitive deficiency may not be as useful as determining what skills (e.g., on some test of academic skills) are deficient under different language conditions. However, the notion that standardized assessment under different stimulus and response conditions is an alternative assessment strategy has been raised and could provide the behavioral assessor useful information in planning an intervention program (cf. Kratochwill, 1977). As Evans and Nelson (1977) observe, "...showing that a child from a different culture fails a test presented in one way (the typical Western European fashion), but passes a similar test presented in another way (using more familiar stimuli), is an assessment of the importance of those stimulus variables for a given task." (p. 640) (see also Cole, Gay, Glick & Sharp, 1971; Price-Williams, 1966; Piersel, Brody, Kratochwill, 1977).

Testing Context. Children develop a set of complex skills which are employed in varying degrees during the administration of a standardized test. Evans and Nelson (1977) note that what can be a major problem for the comparison of test scores across cultures, subcultures, ethnic groups, or social classes can be useful to the professional because the testing situation represents an opportunity to observe the child's style of behavior on cognitive tasks. Nevertheless, they suggest that one major limitation of this procedure is that such observational categories are subjective and frequently no reliability measures are taken. While some scales are specifically developed for this purpose [e.g., Sattler's (1976) Behavior and Attitude Checklist for IQ Testing], the psychologist would need to construct specific scales for different tests.

A second problem is that many of the tasks are not well designed to tap the child's problem-solving strategy. For example, it would be difficult to completely analyze the problem solving activity of a child completing the WISC-R block design without inclusion of covert verbal statements that accompany performance actions.

Thus, it must be stressed that within a behavioral analysis of IQ test performance (and any test generally) is a function of the test situation and a child's interactional history (Bijou, 1976). Within the test situation, performance will be a function of the test items and setting factors in test taking.

Summary and Conclusion

In this chapter we reviewed some alternatives that have been proposed to traditional assessment practices. Traditional assessment practices have usually involved a relatively standardized battery of assessment such as standardized I.Q. and personality measures. A number of alternatives have been proposed, including culture-reduced testing, renorming, adaptive behavior, Piagetian assessment procedures, learning potential assessment, diagnostic clinical teaching, child development observation, and clinical neuropsychological assessment. In each of these areas we found a rather limited amount of empirical research addressing the issue of how these strategies can actually reduce bias in assessment. In some cases there are conceptual and methodological problems in the research. In other areas there is no strong evidence to suggest that any of these things result in better services to children as a function of their inclusion in the assessment process. In order to address this concern, more empirical research taking into account these different alternatives needs to be conducted.

A rather extensive discussion of behavioral assessment strategies was included in the chapter because these techniques have been relatively ignored in the test bias literature. Indeed, there has been a paucity of information relating behavioral assessment techniques to an expanded framework for assessment to reduce bias. After reviewing a conceptual framework for behavioral assessment techniques, we provided a review of specific techniques including, interview, self-report and behavioral checklists and rating scales, self-monitoring, analogue assessment, direct observational assessment, psychophysiological assessment, criterion-referenced assessment, and the use of more traditional assessment procedures within behavioral assessment.

Behavioral assessment methods, on the surface, look like procedures that are relatively useful in expanding of framework for assessment in educational settings with minority and non-minority children. One of the major advantages that some of these procedures have (e.g., direct observational measures) is that they can be used over different phases of the assessment process to determine how effective services are. This feature of behavioral assessment is perhaps one of the stronger characteristics that needs to be emphasized in future assessment work. In addition, criterion-referenced assessment holds great promise for work in this area; however, as noted in this section of the chapter, there are numerous conceptual/methodological features of this form of assessment that have not directly addressed the issue of bias in assessment.

Again, we must conclude that even in the behavioral assessment area where there is strong promise of reducing bias in the assessment process, little empirical research has been conducted on this specific topic. Indeed, the field of behavioral assessment lacks any conceptual framework for dealing with test bias in a systematic way. One of our strong recommendations is that the task of providing a conceptual framework for research must be undertaken in the future.

Finally, although a number of alternatives to traditional assessment have been proposed, at this time it is likely that these procedures should be considered as adjuncts to more traditional assessment until data indicate that any procedure, or a combination of procedures, can make a strong contribution to reducing bias in the assessment process.

Chapter 8

Ethical and Legal Considerations

Assessment and treatment of children's learning and emotional problems necessitates a discussion of ethical and legal considerations especially as they relate to the assessment of culturally and linguistically diverse populations. Specifying that a child has a learning or behavior problem raises issues of labeling and the possibility of professional intervention (e.g., consultation, special education services). Once a judgment has been made that a problem exists, some professional may become involved in attempts to assess and treat the problem. Sometimes, depending on the type of case, a research investigation may also be considered. Any of these procedures involve something intrusive for the child and may expose him/her to a range of risks and possible inconveniences. A child's participation in assessment, treatment, and/or research may involve the following potential intrusive influences: (a) privacy of the child (and parent) may be involved; (b) personal resources (e.g., time, money) may be used; (c) personal autonomy may be sacrificed; (d) the client (and family) may be exposed to physical and/or psychological pain and discomfort; (3) permanent physical and/or psychological damage may occur (Stuart, 1981). Due to these potential negative influences, various guidelines, laws, and moral codes have been developed. In this chapter we review the ethical and legal issues relevant to non-biased assessment of children experiencing learning and behavior problems. These considerations are presented within the context of assessment, treatment, and research with school age populations.

Ethical and Legal Issues: The Context

The assessment procedures reviewed in earlier chapters offer some promise for achieving various therapeutic goals in work with children experiencing learning and behavior problems. Yet, the very fact that these procedures can be used to change feelings, cognitions, and behavior of the child and his/her family raises numerous concerns over the relationship between professional and client(s). Concern for the nature of a therapist-client relationship is not new. Perhaps since the days of Hippocrates individuals have been sensitive to the special nature of the relationship that exists. When one person goes to a designated professional for help, he/she is vulnerable to potential abuse. Over the years, scholars, professional groups, and the courts have raised issues over the nature of assessment and its potential impact on the consumer. Three sources of guidelines (influence) have been established for professionals involved in therapeutic assessment and intervention (Stuart, 1981): (a) law, (b) ethics, and (c) morality. These influences provide a conceptual guide for the professional involved in assessment, intervention and research with minority and nonminority children.

Laws

The future of many assessment and therapeutic programs for children experiencing academic and social disturbances is increasingly being influenced by legal precedents established in the judicial system. Laws provide one of the strongest influences on professional behavior. Generally, laws, whether by statute or case, represent formal principles that govern conduct, action, or procedure. In this sense they can be viewed as guidelines for activities among professionals. In fact, laws establish, in some cases, who can be a professional by judging or who can call themselves a physician, lawyer, psychologist, counselor, and so forth.

Unfortuantely, as Stuart (1981) aptly notes, laws have typically been proscriptive rather than prescriptive. They have typically specified sanctions and penalties for misconduct rather than established guidelines for positive and acceptable activities. Another characteristic of laws is that they have typically been reactive rather than proactive. In most cases laws have come into effect subsequent to some misconduct or misdeed rather than being enacted to prevent various problems, although they may prevent future behavior of the same kind (e.g., discriminatory assessment practices). For these reasons, laws can be regarded as somewhat incomplete guidelines for professionals. Although their formal establishment may represent a potent source of influence, this may be too specialized to provide the kind of information the professional needs in his/her everyday assessment and therapeutic activities.

Ethics and Rights

Ethics are something that nearly every professional agrees we should have, but agreement over the definition and scope of ethics remains a continual source of controversy. Morality is often implied in definitions of ethics as this dictionary definition implies (cited in Krasner, 1976, p. 631):

> "The study of the general nature of morals and of the specific moral choices to be made by the individual in his relationship with others ... The rules or standard governing the conduct of the members of the profession ... Any set of moral principles or values ... The moral quality of a course of action; fitness; propriety (American Heritage Dictionary, p. 450)."

Within this context, the assessor/clinician must make decisions on the basis of what is good and bad for the specific individual, as implied in definitions of morality (Krasner, 1976). In these decisions, the treatment issues of control and prediction of behavior emerge. Braun (1975), for example, raised the following concerns with behavior modification: "Who shall have the power to control behavior?"; "Towards what end shall the controlling power be used?" "How shall the power to control behavior be regulated?" Such issues are, of course, not specific to behavior therapy, and can be seen to emerge in any of the assessment and intervention approaches used for the provision of children's special educational services.

Sometimes a distinction is made between ethics and human rights (Morris & Brown, 1982) For example, if the means used to assess or treat a child with a severe behavior disorder was intolerable to this child (e.g., flooding, punishment, role playing, and s(he) elected not to be involved in the program, a rights question would emerge. In contrast, the professional may become involved in an ethical decision when deciding which assessment/intervention procedure would work with the specific type of problem experienced by the child. Additional concern may arise when assessing minority group children. For example, certain minority group children may feel less comfortable in testing situations (see Chapter 5) and may have their rights violated more easily than nonminority group children.

It is obvious that rights and ethics overlap in practice. Thus, failure to consider a human rights issue would be regarded as unethical. Yet, the manner in which ethical and human rights codes provide guidelines for professional behavior vary and are not often uniform across disciplines (e.g., psychology, special education).

Moral Principles

Moral principles represent an influence on professional behavior inasmuch as they provide guides for conduct that transcend specific laws and ethical codes. Moral principles refer to "...some absolute assumptions about the rights and responsibilities of individuals (Stuart, 1981, p. 717)" As noted above, morality plays a central role in ethical guidelines. No assessment strategy or therapeutic model is free from scrutiny on ethical and moral grounds, but the techniques derived from any particular approach do not imply a particular ethical or moral approach. Bandura (1969, p. 112) raises this issue:

> In discussions of the ethical implications of different modes of achieving personality changes, commentators often mistakenly ascribe a negative morality to behavioral approaches, as though this were inherent in the procedures. Social-learning theory is not a system of ethics; it is a system of scientific principles that can be successfully applied to the attainment of any moral outcome. In actuality, because of their relative efficacy, behavioral approaches hold much greater promise than traditional methods for the advancement of self-determination and the fulfillment of human capabilities. If applied toward the proper ends, social-learning methods can quite effectively support a humanistic morality.

Morality is, of course, the issue that emerges in determining what are proper ends of any assessment or therapeutic procedure.

It should be emphasized that various ethical codes of the professions will not provide guidelines for all issues that emerge in treatment assessment, and research. Even with the best ethical codes, individuals will need to embrace basic moral principles for human conduct. But, basic moral principles vary within and across cultures, and may even be more subjective (and less easily identifiable) sources of guidance and regulation. Presumably, moral thinking is the basis for development of codes of professional conduct, but this has not always been specified.

Issues in Assessment

Virtually all special education programs established for children experiencing learning and behavior disorders involve some type of formal or informal testing and assessment. Although some individuals have made a distinction between testing and assessment (e.g., Mahoney & Ward, 1976; Salvia & Ysseldyke, 1981), we will be using the terms interchangeably (see discussion in Chapter 1). However, as we will see, an important issue that emerges in psychological and educational assessment when legal and ethical issues are embraced relates to the psychometric credibility of the procedures as well as the use for which it is put in making decisions about child intervention. For example, in 1972 it was estimated that more than 250 million tests in the area of academic skills, perceptual and motor functioning, social-emotional functioning, and vocationally oriented skills were administered in education (Hohman & Docter, 1972). In 1975 when Congress passed Public Law 94-142 (20 U.S.C. 1401-1461), The Education for all Handicapped Children Act, large numbers of normal and handicapped children experienced assessment from a variety of school based professionals (e.g., school psychologists, speech therapists, counselors, special educators).

The rapid proliferation of assessment has raised consciousness over the implications this activity has for individuals participating in it. Increasingly, individuals outside the professional community have become quite critical of testing practices and procedures. Three books were quite instrumental in alerting the public to sources of controversy and problems in testing: The tyranny of testing by Banesh Hoffman (1962; The brain watchers by Martin Gross (1962); and They shall not pass by Hillel Black (1963). Concerns over testing have also grown in the fields of psychology and education especially as they relate to assessment and treatment of minority group children (see Chapter 1). Organizations have established formal guidelines for assessment and even statements of policy on proper use of tests (see Chapter 9). A more recent involvement has come from the courts who have been asked to decide on the utility of testing practices in making decisions about services for children and youths. In this section of the chapter, we deal with how these issues influence assessment activities for both minority and nonminority group children.

Criticisms of Assessment

Assessment typically involves some type of relationship between the assessor and the assessee. This relationship may not always be known to the assessee, especially with regard to the potential consequences the information gathered during assessment may have. Numerous criticisms have been advanced against assessment practices and instruments. Among the more common include the allegation that assessment represents an invasion of privacy, assessment may create an unfavorable atmosphere, assessment results in labels, and assessment may be discrimiatory against certain groups.

Invasion of Privacy. The right of privacy is embedded in the U. S. Constitution, but remains remarkably ambigious in some areas of practice. There appears to be two somewhat overlapping aspects that emerge in the privacy concept (Bersoff, 1978). First is the right not to suffer government prohibition as a result of engaging in private activity. The second is the right to be free from government gathering, storage, and dissemination of private information (see Dorsen, Bender, & Neuborne, 1976). Extending this concept beyond the governmental context, Reubhausen and Brim (1965) offer the following definition:

> "The essence of privacy is.." the freedom of the individual to pick and choose for himself the time and circumstances under which, and most importantly, the extent to which, his attitudes, beliefs, behavior and opinions are to be shared with or withheld from others. The right of privacy is, therefore, a positive claim to a status of personal dignity, a claim for freedom ... of a very special kind (pp. 1189-1190).

The essence of the issue in any type of assessment is the right of the person to determine what type of information of a personal nature will be shared with others. Consider the following situation that might be involved in assessing a minority child who is extremely socially withdrawn. A psychologist, believing that data are needed on the peer perspectives on the child might administer a sociometric scale to the child's classmates during regular class sessions. Although cooperation of the school has been obtained, voluntary consent of the students and parents have not been sought.

This situation involves consideration of several issues that are problematic. First of all, informed consent was not obtained of the student (see later discussion of the informed consent principle). Second, the information obtained on the test scale was likely of a highly personal nature and many students may have considered it an invasion of privacy to provide it. Third, it was not specified how the information was to be used. That is, were the data to be disseminated to any personal with identifying information and if so, to whom? Consider further that the psychologist might be asked by the school officials to intervene with other students who are reported to be withdrawn or friendless. The issue of the psychologist offering his/her services to an extremely withdrawn child may emerge. Therefore, the individual's personal privacy may be involved.

Invasion of privacy is then a broad concept that involves several issues, including informed consent, confidentiality, and even psychological stress. Invasion of privacy is more complex when children are involved than when adults are involved. Privacy rights guaranteed by the constitution are granted to adults and generally not to children. Although the courts have granted some privacy rights to children (e.g., Tinker, v. Des Moines School District, 1969), there may be considerable compromise where the child's interest is at stake (Bersoff, 1978).

Privacy issues are also raised in the use of unobtrusive measures in assessment of learning and behavior problems. Usually, unobtrusive measures are taken without the client's awareness so as to avoid sensitizing them (see discussion in Chapter 7). Yet, obtaining such measures may violate the requirements of informed consent and may be perceived as an invasion of privacy. Assessors might consider several alternatives in this area. First of all, for some unobtrusive assessment, the issue of consent may not emerge. Many archival records would be publically available and could be used without any personal identification. For example, the grades a child receives as part of his/her regular evaluation or the number of times the child is sent to the principal's office represent some examples that represent a minimal invasion of privacy. As a second possibility, the child's parents could provide consent for several different types of assessment opportunities, only some of which would be used (Kazdin, 1979). A third option sometimes advanced is to go ahead and conduct the unobtrusive assessments and inform the child/parent subsequently that he/she has the option to have such information remain confidential. Yet, this option may not be acceptable given the possibility that assessment was initially objectable (Kazdin, 1979). Moreover, once the information is obtained, it could conceivably cause a threat to privacy, especially in cases where it has an important bearing on the decision making process or where legal issues are involved (e.g., discrimination, access to special services).

The decision to require students to take tests or examinations (especially those involving personality and/or attitudes) could be done within the context of a panel that considers some of the implications involved. Specifically, the following might be considered:

1. The ability of the test to measure precisely those objectives the school or district intends to measure.
2. The possibility of embarrassing or emotionally damaging children who take the test.
3. The extent to which community mores and values are likely to be affected by the test.
4. The potential benefits of testing.
5. The possibility of using volunteers instead of captive audiences.
6. The steps that will be taken to ensure confidentiality of results, and

7. The possibility of obtaining data without testing (using census reports or public documents, for example) (Sax, 1974, pp. 26-27).

In addition to the above, we would add the need to determine the potential biased nature of the test when used with children from diverse cultural and language backgrounds.

It is recognized that the invasion of privacy issue extends well beyond testing (American Psychological Association, 1981). Many routine activities in our society are aimed at gathering information that is commonly regarded as personal (e.g., opinion polls, or credit card applications). Material provided by students in school as part of regular class activities might also be regarded as such. Yet, it is the responsibility of the professional to follow the legal directives and ethical guidelines advanced that have a bearing on such issues (see later sections in this chapter).

Tests Create an Unfavorable Atmosphere. Another criticism of tests has been that they may create an unfavorable atmosphere for the client or student involved in taking them. Indeed, the requirement that children and youth participate in formal test-taking has created a whole literature on treatment methods to reduce test anxiety. Test anxiety has been a source of some discussion in the professional literature (e.g., Johnson, 1979; Morris & Kratochwill, 1983; Phillips, 1978; Sarason, 1980; Tryon, 1980) and refers to "... an unpleasant feeling or emotional state that has physiological and behavioral concomitants and that is experienced in formal testing or other evaluative situations" (Dusek, 1980, p. 88). It is possible that test anxiety is associated with cognitive and attentional processes that interfere with task performance, although this does not always occur.

Administration of certain individual standardized tests (e.g., IQ) has also been criticized for inducing an unfavorable atmosphere that may result in discriminatory practices to certain ethnic or racial groups (see Kratochwill, Alper & Cancelli, 1980). It is sometimes assumed that if a child does not perform well on individual tests, the results could be an inaccurate reflection of classroom performance (Reschly, 1979). Factors accounting for poor performance on tests may be related to motivational factors or situational anxiety generated by the test and test environmnet (e.g., an unfamiliar situation, examiner). For example, Piersel et al. (1977) found that a pretest vicarious situation in which minority group children viewed a seven-minute videotape of a white examiner testing a minority child under positive conditions (e.g., praise) resulted in only 14.3% of the WISC-Revised (WISC-R) scores being 1 SD below the mean, whereas 42.8% and 52.4% of the scores were 1 SD below the mean under a standard administration and feedback conditions, respectively. Although the findings were discussed within the context of motivational factors, specific anxiety components could also be invoked to explain the performance differences. Children viewing a pretest vicarious interaction may experience reduced levels of anxiety that then have a positive influence on performance.

Of course, deliberate attempts to create anxiety or fear among children during testing situations may prove unethical. Yet, the administration of tests to children is a commonplace event in the school and community. A major issue here is that efforts should be made to reduce the severe negative influences that accompany testing. A rather large body of literature suggests some useful intervention procedures for this problem (see, for example, Tyron, 1980). In the area of assessing children's fears and phobias, the psychologist must consider that administering some particular assessment device could have a negative impact on the client. Two considerations emerge from this type of assessment. First, attempts should be made to reduce any negative emotional aspects that surround the assessment procedure. This would be in accord with sound ethical practice where stress should be minimized. Second, the assessor must consider that any anxiety created by the testing itself may lead to inaccurate results and hence could possibly lead to misguided intervention procedures .

Assessment Results in Labels. A major objection to assessment is that it frequently results in labeling the child in a way that may prove destructive. Concern over the labeling process has been extensively discussed in the professional literature (e.g., Gordon, 1975; Guskin, 1974; Hobbs, 1975a, 1975b, 1975c; MacMillan & Meyers, 1979; MacMillan, Jones, & Aloiu, 1974; Mash & Terdal, 1981; Mercer, 1973, 1975; Ross, 1980; Rowitz, 1974). Yet, labeling children may have a number of positive features (Rains, Kitsure, Duster, & Friedson, 1975). First of all, labels may help summarize and order observations which in turn help professionals communicate. For example, professionals with diverse backgrounds can talk about "organic mental disorders" (DSM-III) and have some general understanding of what is involved in the problem. Second, labels may in some cases facilitate treatment strategies for a particular disorder. Given that a learning disorder can be reliability diagnosed, several of the available treatment approaches described in the professional literature could be employed. Third, labels may serve as an organizer for scientific research (e.g., epidemiological, etiological, and treatment) on a particular disorder. Fourth, labels may serve as a reference point for tolerance or acceptability of childhood behavior (Algozzine, Mercer, & Countermine, 1977).

More often, negative features of the labeling process have been raised. Concern over labeling became especially acute during the late 1960s and early 1970s as legislation emerged dealing with this issue. The perceived negative by-products of labeling have been of particular interest when discussing the disproportional representation of minority group children in educaational diagnostic categories. In addition to general concerns over diagnostic classification systems, increased attention has been focused on how children are labeled in the schools. Such growing attention has been due in part to (a) the need to classify students for certain purposes (special services) and to assign names to these classifications, (b) the non average characteristics of certain groups (emotionally disturbed), and (c) the propensity to associate children with the name of the group they have been assigned to (MacMillan & Meyers, 1979).

When a decision is made to assess a child experiencing a learning or behavior disorder, several potential issues emerge in the process. First of all, a possible concern of clients and their caretakers relates to the possible label, diagnosis, or classification that may ensue. In such cases it is not so much the labeling process itself as it is the potentially negative influences associated with the label (Pruch, Engel, & Morse, 1975; Smith, 1981). In our culture the use of some formal label may frequently be associated with the assignment of a negative value such as "sick", "disturbed", "mental", and so forth. These values may further cause emotional suffering on the part of the child and/or parents. Second, beyond the specific concerns with labeling per se, there may be long term negative consequences associated with the labeling process such as lack of a regular education, denial of employment, among others. Third, the use of formal diagnostic classification systems (e.g., DSM-III, PL 94-142) may lead to a violation of human rights in that the labels employed may be shared with others without the informed consent of the client (Smith, 1981). Smith (1981) notes that clients are not usually informed that when they sign confidential information release forms they are often giving blind consent to release specific diagnostic classification data. Of course, this could occur for both the parent and child. Smith has argued that the APA should amend the ethical guidelines for provision of psychological services to take into account this apsect of practice.

Assessment of children experiencing learning and emotional problems in school settings may lead to special class placement as well as labeling (e.g."emotionally disturbed"). Questions can be raised over the efficacy of this process and some empirical work has appeared in the field of psychology and education on this issue. The primary concern here is whether or not labels such as "emotionally disturbed" have an adverse effect on the child in the sense that labels may bias the professional toward seeing more pathology or deviance than otherwise would have been perceived without the label. While some writers have noted the potential negative effects of labels, (e.g., Catterall, 1972; Reynolds & Barlow, 1972), there is no uniform evidence for this negative influence. As Mash and Terdal (1981) observed, some research has shown that a particular child behavior, when believed to be exhibited by a "disturbed child", may produce different reactions than when believed to be exhibited by a 'nondisturbed or normal child" (e.g. Stevens-Long, 1973). Also, other research has shown that observers may tend to overestimate negative behaviors in a group of children labeled behaviorally disturbed whereas they underestimate negative behavior in a group labeled normal (Yates, Klein, & Haven, 1978). Several reviews of the literature have not found support for the adverse effect of labeling (e.g., MacMillan & Meyers, 1979; MacMillan et al., 1974; Guskin, Bartel, & MacMillan, 1975; Kratochwill et al., 1980; Reschly, 1979). Some discrepancies in this area may be an artifact of the methodologies employed (Reschly, 1979). As an example, in studies where college students or teachers are provided only the label and/or not or only brief exposure to the labeled child, a relatively large expectancy effect is found (Ysseldyke & Foster, 1978). Yet, in studies employing the same basic methodology but a more lengthy exposure to the labeled child, the expectancy effect is either diminished over time or is not found (Reschly & Lamprecht, in press; Yoshida & Meyers, 1975). Moreover, the fact that many studies have not been carried out in the clinical setting threatens the external validity of this empirical work (Kratochwill, et al., 1980).

Concerns have also been raised over a possible "self-fulfilling prophecy" effect. The issue here would be whether or not a child labeled as disordered will be perceived in a negative manner, thereby contributing further to the problem. Research in this area has not clarified the issue. For example, Rosenthal and Jacobson's (1968) work has been criticized for methodological inadequacies (e.g.,, Elashoff & Snow, 1971; Humphreys & Stubbs, 1977; MacMillan et al., 1974; Snow, 1969; Thorndyke, 1968) and at least some research with the emotionally disturbed label (e.g., Foster, Ysseldyke, & Reese, 1975) has not supported the self-fulfilling prophecy notion.

Conceptual problems with the potential negative impact of labeling have also been raised. The label itself may not be the sole cause of negative experiences that are presumed to be associated with it. Thus prelabeling behaviors exhibited by the child that led to labeling, as well as consequences associated with the label may account for the negative influences. As Mash and Terdal (1981) note, the informal labeling process and interpretations of formal labels by various individuals (parents, teachers, etc.) may have the greatest impact. Once a child has been labeled emotionally disturbed (e.g., emotionally handicapped) it is conceptually impossible to attribute the negative influences to the formal labeling set (MacMillan & Meyers, 1979). Thus, the labeling process appears to be conceptually complex and any variance attributed to the labeling experience must take into account several factors noted by MacMillan et al. (1974):

1. Prelabeling experiences.
2. The effect of the label versus the pereceptions of the services received by the individual once labeled.
3. The effects of formal versus informal labels on the measure of interest as well as the agency and individuals who append the label.
4. Cases where children carry multiple labels simultaneously, such as the delinquent-EMR, or disadvantaged-LD.
5. The response of the child and the family to the appended label; most noticeably whether they deny or accept its validity or whether the salient attribute is highly valued by the child's subcultural group (MacMillan & Meyers, 1979, pp. 179-180).

It appears clear that future research will be important in sorting out the conceptual and methodological issues in this area. It does appear that assessment will continue to result in diagnosis, classification, and labeling. As Achenbach (1974) has noted, "The basic question is not whether to classify, but how to classify" (p. 543). Many of the issues in the literature have been advanced toward traditional assessment and diagnostic systems. Although behavioral assessment sometimes leads to formal diagnosis and labeling (see Chapter 3), it is unclear what influence the emerging behavioral assessment/classification schemes will have on children. Some evidence suggests that behavior therapists may be less easily biased by labels (e.g., Langer & Abelson, 1974), but much work remains to be done on this issue. When research is conducted, it will be productive to consider the already developed conceptual and methodological issues advanced in this area.

Assessment Results in Discriminatory Practices. A central criticism of assessment is that it leads to practices that are believed to be discriminatory against certain individuals or groups, usually minority racial and ethnic groups (Alley & Foster, 1978; Flaugher, 1978; Kratochwill et al., 1980; Oakland & Matuszek, 1977; Reschly, 1979; Sattler, 1974). Minority children frequently face the problem of misclassification in school systems with black children being three times as likely as white children to be placed in classes for the educable mentally retarded (see SCF [1980]; and the report by the Education Advocates Coalition on Federal Compliance Activities to the Education for All Handicapped Children Act ([Public Law 94-142, April 16, 1980]). The concept of nondiscriminatory or non-biased assessment has been a central theme in recent federal legislative and judicial actions that have provided guidelines for assessment practices.

The concept of nondiscriminatory assessment has invoked two primary legal, ethical, and moral issues, namely assessment of minorities and the use of certain traditional assessment devices and procedures (e.g., IQ test) in this testing process. As noted in Chapter 1, traditional assessment procedures have been the primary focal point of criticism with several alternatives examined. Among the more common recommendations in the area of child assessment that we discussed in Chapter 7 has been the call for a moratorium on conventional tests, elimination of speical class placement, language translation in testing, use of minority group examiners, modification in test procedures (e.g., providing reinforcement), and creation of so called "culture fair" tests. Although it is beyond the scope of this chapter to review each of these proposed alternatives within an ethical-legal context, it should be noted that each provides numerous conceptual and methodologcial problems.

Perhaps the major limitation in work in the field of nondiscriminatory assessment has been the conceptualization of discrimination within the context of minorities and traditional testing. As an alternative, assessment procedures should be evaluated on dimensions of discrimination within the context of how they influence children, regardless of race or cultural background (Kratochwill et al., 1980; Reschly, 1979). Within this proposal is the thesis that assessment that results in equally effective interventions for both minority and nonminority group children has met the spirit of being nondiscriminatory. When assessment practices result in differentially ineffective services across groups, the possibility of discriminatory practices must be considered.

Legislative and Judicial Influences

Psychology and education are becoming increasingly regulated. This is especially true in the area of assessment activities occurring in these fields. At one time, the courts were not involved in examining the assessment activities of psychologists and educators. One reason for this stance was that the courts pleaded lack of expert knowledge (Bersoff, 1981). But this is definitely changing as reflected in decisions of the Supreme Court, lower federal and state courts, as well as in Congress and in federal administrative agencies. For example, the Supreme Court has been involved in such activities as the influence of compulsory education laws, the requirements of due process prior to application of disciplinary and academic sanctions, and the allocation of financial resources to "poor" schools. The lower federal and state courts have rendered decisions on such areas as the right to education for Handicapped pupils, appropriate identification of learning disabled children, and the right of schools to expel disruptive handicapped students, and assessment of minority group children. Congress and federal administrative agencies have been active inasmuch as in 1964, the Civil Rights Act passed by Congress contained antidiscrimination provisions guarding against discrimination based on race, color, or national origin. These were followed by the passing of the Rehabilitation Act of 1973, the Family Education Rights and Privacy Act, and the Education for all handicapped Children Act of 1975 (Public Law 19-142).

The reason that the courts have demonstrated a willingness to render judgments on assessment issues relates to the degree to which the constitutional rights of an individual have allegedly been violated.

Specifically, in the assessment area the three constitutional principles of equal protection, due process, and privacy, have been the focus of intervention by the courts (Bersoff, 1981). For example, in the case of assessment, the right of equal protection has been interpreted (in part) as the right to an equal educational opportunity and has been used successfully in some cases (e.g., Mills v. Board of Education of the District of Columbia, 1972; Penn. Association for Retarded Children v. Commonwealth of Penn., 1972). For example, a severely disturbed child would be entitled to a public school education despite the handicapping condition. In the case of due process, the fourteenth amendment requires individual notification in a fair and impartial manner where interests protected by the Constitution are either restricted or rescinded. Specifically, the due process clause applies where the individual's interest in life, liberty, or property are being considered. For example, a school cannot label a child "emotionally disturbed" unless there is a formal hearing conducted. Thus, the potentially negative consequences of such labeling must be considered. In the case of the right of privacy, the judiciary has not defined the concept, but has indicated activities within its scope. Generally, the concept has been broadened to include freedom from unreasonable intrusion into family life by individuals providing mental health services (Bersoff, 1981).

As a consequence of legislative and judicial influences in assessment practices, several procedural requirements have been established. For example, the regulations of PL 94-142 represent requirements for the evaluation of children who might be demonstrating learning and behavioral disorders and who are being considered for special education services in a public school setting.

Some specific issues that would create a procedural concern for a psychologist or other mental health professional practicing in a public school setting involve notice, consent, and access to records (Bersoff, 1981; Martin, 1979).

Notice. One of the first procedural safeguards that must be met is formal notice. Such notice cannot be incomprehensible or intimidating and cannot come after the fact. Public Law 94-142 (45CFR 121a. 504(a) (1) - (2) requires written notice "a reasonable time before the public agency proposes to initiate or change (or refuses to initate or change) and the identification, evaluation or placement of the child or the provision of a free appropriate public education to the child". For example, in a typical case of a child being considered for special services due to severe anxiety, a psychologist must inform the parents at each step of the process, including assessment. The school would notify the parents that there may be a problem and that a professional evaluation will be conducted on the child. Once the evaluation is accomplished (given that consent was obtained), the psychologist would need to inform the parents what will be done next (e.g., an interventon program), or that nothing will be done. Notice would, of course, extend throughout the intervention phases as well.

Notice is not always satisfied easily. It must be written in a manner that is comprehensible to the parent or guardian. If the parent cannot read or only reads a foreign language, other means must be employed to make notice understandable. The specific action proposed for PL 94-142 is:

1. The proposed action must be stated.
2. There must be an explanation why the school proposes the action.
3. A description must be given of the alternatives which were

considered before the proposed action was decided on.

4. The reasons must be explained why the other alternatives were rejected.

5. Each evaluation procedure, test, record, or report that the agency will rely on as a basis for the proposed action must be described.

6. Any other factors relevant to the agency's proposed action must be described [45CFR 1212.505(a)(1)-(4)].

Consent. Notice does not imply consent. Technically, notification refers to supplying information about impending actions whereas consent requires affirmative permission before actions can be taken (Bersoff, 1981). Within the context of PL 94-142, consent is required for only four things (Martin, 1979, p. 102):

1. The initial evaluation of the child.
2. The initial placement of the child.
3. Evaluation before a "subsequent significant change in placement".
4. Before release of records to persons not already authorized to see them.

Like notice, informed consent is sometimes difficult to define in any technical sense. For example, the question of who can give consent remains controversial. Although in the case of children one might expect that the parent would be the typical individual to render consent, open opposition from the child may cloud the issue (Martin, 1979). In Bartly V. Kemens and J. L. v. Parham two federal courts required an opportunity for hearing when parents consented to place youths, over their protests, in institutions. Generally, it is recognized that informed consent contains three basic characteristics that must be upheld to meet the spirit of the concept: Knowlege, voluntariness, and capacity. Each of these issues is discussed in more detail later in the chapter. Essentially, the three concepts are applicable to informed consent in assessment, treatment, and research.

Access to Assessment Records. Psychologists and other professionals engaged in assessment activities with children experiencing learning and behavior problems typically (and hopefully) generate a considerable amount of data. A question that occurs in this activity is whether or not the data are accessible under existing law. In the past, psychological assessment data, test protocols, and client responses have been guarded to prevent public disclosure. However, much assessment data are now available to the public, due to the Family Education Rights and Privacy Act (1975). This act sometimes called the"Buckley Amendment", has been incorporated, in part, in PL 94-142. Any educational institution receiving federal funds by the U. S. Office of Education must allow parents access to the records maintained on their child and the right to challenge any information believed to be inaccurate or damaging to the child. In school settings an educational record refers to records that are directly related to a student and are maintained by the educational agency or by a party acting for the agency (45 CFR 99.3).

A question that immediately arises here is whether assessment information and test protocols are accessible under existing laws. At this time, the issue is not clear because no cases have ruled to clarify this point (Bersoff, 1981). One of the key issues is whether or not the psychologist, psychiatrist, social worker or other professional reveals information to individuals in the course of providing input on a case. For example, a psychologist might administer a rating scale or checklist to a child as part of an assessment to determine if special class placement is necessary. If responses were revealed in a team meeting to determine class placement, such information would be considered part of the educational/assessment record. Presumably, this would also be true in cases where a psychologist from an outside agency (e.g., mental health clinic) is providing input in the case. Although the issues are far from settled, analogous cases in industry (e.g., *National Labor Relations Board v. Detroit Edison*)[13] suggest that parents will increasingly be granted access to psychological assessment material. In *Lora v. Bd. of Education of the City of New York (1978)*, the court noted that a failure to provide parents with "clinical records" from which placement decisions are made does not follow due process. However, the failure to carefully define clinical records still makes the issue ambiguous (Bersoff, 1981).

Bias in IQ Testing

In addition to guidelines promulgated by court, legislative, and governmental agencies regarding issues related to procedural concerns such as consent and access to records, recent court rulings have specifically addressed the issue of bias in the tests used by schools in the classification and placement of minority group children in classes for the mentally handicapped.

While not directly addressing assessment techniques, two early cases, *Pennsylvania Association for Retarded Children v. Commonwealth of Pennsylvania* (1971) and *Mills v. Board of Education of the District of Columbia* (1972), had a definite impact on assessment in the schools. Both of these cases were primarily concerned with the rights of retarded children to a free public education. The decisions in favor of the plaintiffs indicated that tests could not be used to exclude children from school as uneducable.

The following year in Louisiana, a ruling in Lebanks v. Spears (1973) regarding the exclusion of children from public school stated that districts could not exclude children from public school who it decided were uneducable, and also set some general guiudelines for placement in classes for the mentally retarded. The ruling stated that such placement can only take place with evidence indicating an IQ below 70 obtained from an individually administered intelligence test and subnormal adaptive behavior. In addition, the court ruled that neither measure could be inappropriately influenced by the sociocultural background of the child.

In another case not directly addressing assessment, Lou v. Nichols (1974), a California school district's was charged with not providing adequate language instruction to all Chinese-speaking students. The ruling of the court, however, directly addressed the districts assessment practices by ordering a task force be set up in the district to insure that bilingual and non-English-speaking children were properly assessed. The courts first directly addressed the charge that psychological tests were biased against black children in Hobson v. Hansen (1967). In this case the plantiffs charged that a tracking system in the District of Columbia public schools was discriminatory in that it led to an overrepresentation. The court decided in favor of the plaintiffs, claiming that the standardized group tests employed in decision making focused on academic achievement. The court concluded that in order to track, it was necessary to assess the children's capacity to learn, something not assessed through the standardized aptitude tests used by the district. The Hobson decision ended the era of group ability testing for classification purpose and led the way for a series of cases that directly addressed the issue of bias in psychological testing.

Two cases that followed Halron in the 1970's brought attention to the potential bias in individually administered intelligence tests in assessing children whose primary language in other than English. In Diana v. California State Board of Education (1970) the plantiffs charged that a disproportunate number of Hispanic children were placed in EMR classes on the basis of IQ tests that they claimed were unfair to bilingual children. Evidence was offered that the nine plantiff's IQ scores were found to be on the average 15 points higher when tested by a bilingual examiner. The out of court settlement resulted in an agreement with the state that all future testing of non-Anglo-American children be conducted (1) in both English and the child's primary language, (2) with tests that were not dependent on unfair verbal questions or vocabulary, (3) by certified school psychologists, and (4) with an assessment battery that was multifaceted, including educational, developmental and adaptive behavior measures as well as intelligence tests. It was further agreed that all Mexican-American and Chinese-American children who were in EMR classes at that time would be reevaluated under the new guidelines and that any district that continued to have a disproportionate percentage of bilingual children in EMR classes would have to provide the State with an explanation for the disparity.

The second case also addressed bias in individually administered intelligence tests when assessing children whose primary language is other than English. In this case, Guadalupe v. Tempe Elementary School District, (1971), the plaintiffs were Arizona Mexican-American and Yaqui Indian children. A settlement similar to Diana was arranged out of court.

While Diana and Guadalupe focused on bias in using individually administered intelligence tests with bilingual children, two additional cases, *Larry P. v. Riles*, (19172, 1974, 1979) in California and *PASE v. Hannon* in Illinois (1980), focused on the alleged bias of individually administered IQ tests in assessing black children. The plaintiffs in both cases charged that, as a result of the use of individually administered IQ tests, black children were being disproportionally placed in EMR classes. Evidence was offered in both cases that re-evaluation of the plaintiffs with examiners sensitive to the culture from which the child came, produced IQ scores that disallowed identification and placement in EMR classes. In the *Larry P.* case, for example, the plaintiffs were readministered the identical IQ test that lend to their EMR classification, but only after rapport was established between the child and examiner, and the setting was reduced of distractions. In addition, some items were reworded and the children's responses were evaluated in the context of what were considered by the examiners to be intelligent approaches to solving the problems posed in the items. Despite the similarity of issues and complaints (i.e., alleged violation of both constitutional and statutory law) Judge Peckham in the *Larry P.* case and Judge Grady in the *PASE* case reached different conclusions. Judge Peckham ruled in favor of the plaintiffs and Judge Grady ruled in favor of the defendents.[4]

In the *Larry P.* case Judge Peckham found that the California State Board of Education violated both the constitutional rights of the plaintiffs under the "equal protection" clause of the Constitution and statutory laws embodied in the Civil Rights Act of 1964, section 504 of the Rehabilitation Act, and Public Law 94-142. With respect to the latter, which required the State to demonstrate the reasonableness of a system that resulted in discriminatory effects (i.e., disproportionate representation), Judge Peckham found for the plaintiffs on the grounds that (1) the tests were culturally biased and (2) there was no demonstrated relationship between black children's IQ scores and school grades. After listening to the testimony of expert witnesses, Judge Peckham concluded that there was not sufficient evidence to support as legitimate the average differences found between white and black children's performances on IQ tests. He rejected the arguments that the differences were the result of either genetic or environmental factors and consequently concluded that the differences must then be a function of bias in the tests. Judge Peckham also concluded that there was inadequate validation of the test for use with black children as called for under PL 94-142. Evidence offered that showed correlation between IQ tests and standardized tests of academic achievement were judged inadequate because Judge Peckham perceived a lack of difference between the measures. Disregarding these findings, Judge Peckham was left with little validity evidence to judge reasonable the State's use of the test in labeling and placing black children in EMR classes.

Judge Peckham also found that the State had violated the right of the plaintiffs under the "equal protection clause" of the Constitution. To find the state in violation of this clause the plaintiffs were required to show that it was the intent of the State to discriminate against the plaintiffs. Judge Peckham interpreted intent to mean that the State willfully engaged in a process that it knew would result in disproportionate representation in EMR classes, classes he labeled a inferior and stigmatizing. Judge Peckham found that the impact of the State Department's action with regard to using IQ tests was not only "foreseeable" but "foreseen" thus allowing for the judgment of intent on the part of the State.

The ruling of Judge Peckham made permanent an injunction ordered in 1981 banning the use of standardized intelligence tests in identifying black children for EMR classes[14]. If such tests are to be considered in the future the State will have to seek the approval of the court. Approval will be granted on condition that the recommended test is empirically supported as valid for placing black children, not racially or culturally discriminatory, and capable of being administered in a nondiscriminatory manner. In addition, the State was required to re-evaluate all black children currently in EMR classes who had been placed using standardized IQ tests and design individual educational plans for all black students returning to regular classrooms. Finally, the State is required, to date, to demonstrate the effectiveness of district plans to correct the proportional imbalance of black children in EMR classes.

Within a year after Judge Peckham's decision in Larry P., Judge Grady rendered her decision in PASE. In finding a favor of the defendent, Judge Grady agreed with Judge Peckham that EMR placement was inferior to regular class placement, but he concluded that there was insufficient evidence that IQ tests were biased. Judge Grady was unempressed by expert testimony and decided in an extraordinary move to judge for himself the validity of the WISC, WISC-R, and Stanford-Binet. Judge Grady read all questions and answers from these tests to the court, determined for himself that a total of nine items were biased, eight items from the WISC and WISC-R and one item from the Stanford-Binet. Judge Grady concluded that these items were insufficient to appreciably influence classification and placement decisions, especially when considered among the additional mandated supporting data for placing children in EMR classes. While he accepted the argument that any one measure could result in bias decisions, he judged the entire placement process identified in PL 94-142 to be a sufficient failsafe system to protect against discrimination. After discounting test bias as the reason for the discrepency between the average IQ scores of black and white children, Judge Grady concluded the difference was caused by socio-economic factors, indicating that poverty was the culprit for the lower average IQ scores of black children.

Bersoff (in press) identifies several similarities and differences in his critique of the Larry P. and PASE cases. Both courts similarly judge special education to programs to be inferior to regular education programs. Judge Peckham in her decision labeled them "dead-end", "isolating", "stigmatizing", "inferior", "substandard", and "educational anacronisms." Judge Grady identified inappropriate placement in an EMR class as an "educational tragedy" and "totally harmful". In addition, both Judges agreed that a multifaceted assessment is necessary for proper classification and placement. However, Judge Grady saw the present process as effective in addressing concerns of bias while Judge Peckham concluded that, despite the mandate for a multifacited assessment, in practice, the IQ score has a disproportionately large impact on EMR decision making.

Bersoff (in press) identifies the major disagreement in the rulings as their difference in perceptions of the biased nature of IQ tests. Judge Peckham concluded that IQ difference across races was an artifact of testing while Judge Grady eluded that the differences were real. Bersoff concludes that despite which decision one favors, both must be considered inadequate given the basis on which they were made. Indeed, Bersoff (in press) in speaking of Judge Grady's decision, concluded that " [T]he method by which he reached that judgment was embarassingly devoid of intellectual integrity" (p. 88).

Judge Peckham's decision that there should be no difference between the average black and white IQ scores is premature to say the least. As reviewed in previous chapters, evidence to date does not support this conclusion. The ramification of this decision, that is that only IQ tests that do not show average racial differences can be used in the future, bears the potential of being more harmful to educationally needy black students than the system now in place. Also, this decision may ultimately lead to tests with less predictive validity than those currently employed.

Judge Grady's attempt to subjectively determine the content bias of IQ test items has already by shown in the literature to be a ineffective procedure (see Chapter 4). His folksy method of judging cultural bias seriously calls into question the value of his conclusions.

When we evaluate the reasoning evidenced in the decisions of Judges Peckham and Grady within the concepts of bias offered in the present review we can better understand how such differences can result. From our analysis, it appears that Judge Peckham adopted an egaliterian definition of test bias by accepting the assusmption that there should be no IQ differences among races. This adoption is in deference to more technical definitions as found in our review of internal and external construct bias. In addition, Judge Peckham appears to have weighed heavily those aspects of the case which suggest his concern for what we have termed outcome bias. As mentioned above, Judge Peckham discounted those predictive validity studies that use nonsituation specific criterion measures (i.e., standardized achievement tests) and that bear more heavily on establishing the valildity of the construct in favor of criterion measures that reflect critical outcomes (i.e., school grades. Emphasis on school grades and demonstration that IQ tests employed in the future show empirical evidence of being able to predict whether or not a child can be successful in a regular class setting with remedial help points to this conclusion. This is a less strigent criteria, however, than we offer to give evidence of outcome validity. From our perspective, the outcome validity of a test employed in decision making for intervention purposes should require evidence of the potential effectiveness of the intervention, with outcome bias existing when the assessment yielded interventions with differential effectiveness across groups. Judge Peckham has only required that one be able to give evidence of the potential failure of an in-class intervention. By his definition, a child could still wind up by default in special education. Such a definition also fails to acknowledge the vital link between assessment procedures and the intervention that follows.

Judge Grady, on the other hand, did not appear to be influenced by issues of outcome bias. Rather, his focus appeared to be on internal construct bias. Consistent with the thinking of those who initially addressed the issue of cultural bias in tests (see Chapter 4), Judge Grady turned to an examination of the content of the tests and his subjective appraisal of bias. The items he ultimately identified as biased items in this manner have not been empirically demonstrated as biased.

The final resolution of these issues will be settled in higher courts. Riles has appealed his case to the Appleate Court and chances are that the Supreme Court will ultimately rule on the case. Professionals in the area of assessment will hopefully have sufficiently cleared up their own understanding of the issues to provide adequate direction to the Court.

Issues in Intervention

As we have noted in the previous section, various ethical and legal issues have been raised in the assessment of children. However, the assessment of children frequently (and hopefully) leads to an intervention and this raises further ethical and legal concerns of similar magnitude. Intervention with children frequently leads to outcomes that extend beyond the clinical behaviors for which treatment was focused. A child's life may be changed dramatically as a function of participation in psychological or educational treatment.

The ethical and legal issues discussed in this section apply to all the therapeutic procedures that would usually be used for children experiencing learning or behavior problems in education settings. However, even though writers from diverse orientations have provided discussion of ethical and legal issues (e.g., Koocher, 1976; Szasz, 1965), various writers have suggested that some approaches deserve special consideration. For example, special concerns have been raised over psychoanalytic therapy for failing to be of demonstrated efficacy in treatment of neurosis (Wolpe, 1981). Wolpe (1981) has been particularly critical of psychoanalysis to demonstrate improvement, even after many years of treatment.

> To keep patients interminably in therapy is an immoral practice and a social blot on the psychological profession. We are all tainted by it. Perhaps in years gone by, one could have argued that there was nothing better to offer and that the still-suffering patient at least had the benefit of support. But it is a moral requirement of any health professional to know art in his or her field and be able to offer patients alternatives when the methods used have failed (Wolpe, 1981, p. 163).

In addition, special concerns have also been raised over behavior therapy. Some of these issues have been prompted by "false images" in the lay field [e.g., books such as Mitford's (1973) Kind and usual punishment: The prison business, newspapers, and films such as Clockwork Orange] as well as in the professional community (i.e., inaccurate descriptions of behavior therapy) (Wolpe, 1981). Yet, the issues extend beyond this feature. For example, Ross (1980) noted that ethical issues are particularly critical for behavior therapists because (1) behavior therapy is a very effective method for bringing about behavior change, and (2) the rudiments of behavior therapy are relatively easy to acquire so that individuals other than well-trained behavior therapists can use and hence, misuse them (p. 62). Others, such as Friedman (1975) have also noted that behavior therapy poses special issues and problems not raised by other therapies and therefore this approach requires special regulation. Issues that have been raised include the view that the basic value premises of behavior therapy may be antithetical to freedom and personal growth (e.g., Winett & Winkler, 1972), the view that behavior therapies include therapeutic assumptions that will led to poor therapeutic results (e.g., Arieti, 1974), and the view that behavior therapy provides a special form of control over others (e.g., Pines, 1973; Szasz, 1975). Concerns have also been raised in the media about the application of behavior therapy in schools, prisons, and in society at large.

Despite concerns raised over behavior therapy, the American Psychological Association Commission on Behavior Modification (Stoltz and Associates, 1978) adopted the following position:

> The commission takes the position that it would be unwise for the American Psychological Association to enunciate guidelines for the practice of behavior modification. The procedures of

> behavior modification appear to be no more or less subject to abuse and no more or less in need of ethical regulation than intervention procedures derived from any other set of principles and called by other terms (p. 104).

The commission went on to stress that regulation of behavior modification to the exclusion of other therapeutic procedures would possibly lead to the demise of the practice of behavior modification in those settings to which guidelines apply. As Goldiamond (1975, 1976) has argued, with special guidelines for behavior modification individuals may be prone to use administratively simpler procedures (i.e., those with little or no annoyance, delay or cost) that may be less effective than behavioral techniques. Moreover, it was noted that specific prescriptive and proscriptive guidelines could curtail developments within the field (see Agras, 1973). Thus, all psychological interventions were said to embrace the same ethical issues that behavior therapy embraces. In this regard, a primary recommendation of this commission was that individuals engaged in psychological interventions subscribe to the ethics and guidelines of their professions. We certainly concur with this perspective and would note that individuals engaged in psychological interventions for children follow the guidelines of their respective professions.

Typically, mental health and educational professionals belong to more than one professional organization that provides guidelines. For example, a psychologist might belong to APA but also subscribe to the guidelines of an organization in his/her area of expertise (e.g., education, mental retardation, behavior therapy). In this regard, the professional organization may offer a rather detailed list of guidelines for theraputic intervention. Such is the case with the Association for

Advancement of Behavior Therapy who (AABT) offered the ethical issues for human services. These guidelines are to be considered prior to implementing a behavior therapy program. Technically, however, the guidelines are not related specifically to behavior therapy because each issue could be considered by a therapist implementing any type of intervention.

In the following section of the chapter we review some ethical and legal issues that must be considered in implementation of interventions for children experiencing learning and behavior problems. These issues include control of behavior, agents of control, informed consent, selection of intervention, monitoring intervention, and therapist qualifications.

Control of Behavior

The intervention procedures used in educational settings raise ethical and sometimes legal issues over the control of behavior. Behavior here refers to thoughts, feelings, images, as well as overt behavior. Control of behavior refers to exerting some kind of power over people by manipulating the environment to increase or decrease behaviors (Ulrich, 1967). A major concern here is that behavior will be manipulated toward some undesirable ends (Kazdin, 1980; London, 1969). Individuals disagree as to which type of goals or ends are desirable and so there is usually controversy over this issue. As is evident from a review of interventions in contemporary journals in psychology and education, these extend well beyond psychological procedures. For example, technological and biochemical interventions have been and are being used to treat children who have learning and behavior problems.

Issues of behavior control are especially a major concern in intervention with children. Generally, children can be "controlled" more easily than adults. Moreover, from an early age the child is totally dependent on others (Hobbs, 1975). The point at which the child is able to control his/her own behavior has been the source of much controversy and is certainly far from settled (Melton, 1981). The central issues relating to ethics and behavior control with children experiencing learning and behavior problems center around two major issues:[15] (1) the issue of controlling children who have difficulty gaining counter control over their own environment (e.g., the child is unable to learn in the regular class), (2) the belief among mental health professionals that their interventions are being implemented in "the best interests" of a child to assist him/her to develop satisfactory adjustment. The ethical issue here refers to the use of some intervention procedures used in treatment, whether psychological or psychopharmicological which may be essentially ethically neutral; they have the potential for use or misuse. Thus, within the context of any intervention program, it may not be so much the intervention, but rather the manner in which the intervention technology is used by the professional that raises the ethical issues.

On the other hand, there may be some treatments that could be regarded as ethically troublesome. For example, serious ethical and humanitarian issues have been raised in the use of implosion/flooding treatments with children (Graziano, 1975; Graziano, DeGiovanni, & Garcia, 1979). Such a procedure is usually quite aversive to the client, and yet the child may not feel free to withdraw from treatment. Moreover, as noted by some writers (e.g., Graziano et al., 1979; Ullman & Krasner, 1975) implementation of implosion requires considerable skill so that the aversive treatment procedure is associated with the feared stimulus rather than the therapist.

The intervention procedures used in educational setting do not always specify how various treatment goals might be attained. A behavior therapist might recommend a modeling treatment strategy for a hyperactive child. These procedures provide a therapeutic technique to reach some goal - reduction of activity level. Yet, the goal of activity reduction is a value judgment on the part of the therapist and perhaps for society at large. Also, individuals may not agree on societal goals. Nevertheless, as Kazdin (1980) notes, "A scientist might well be able to predict where a preselected goal will lead, make recommendations to avert undesirable consequences, or investigate the actual effects of pursuing certain goals. Yet, the initial selection of the goal is out of the scientist's hands." (p. 311). This is frequently the case for psychologists and other mental health professionals who work in institutional settings. For example, the school psychologist would be expected to work toward the goal of getting the emotionally disturbed child back into the regular classroom. This control issue may conflict with both the child and the parents' goal.

The fact that many goals are out of the scientist/professional's hands does not deny the possibility that they can reshape or change the goal. Some goals such as the one indicating that all children should attend school might be modified in the individual case. Generally, however, the goals of the client are compatible with social goals. This is usually true in cases of learning and behavioral problems where unpleasant and even aversive consequences are associated with the problem.

Although the many techniques used to treat children are ethically neutral, there will be no neutral or value-free position in actual implementation of the technology (Kazdin, 1980). Thus, any practice and study of human behavior change does not remain value free (Krasner, 1966; Rogers & Skinner, 1956). Thus, endorsing the goal of the individual, socialization agents, the institution, or society reflects a definite value position (Kazdin, 1980). Many of the intervention procedures described in the professional literature provide useful techniques for children experiencing learning and behavior problems. The issue of who may control these techniques and establish the goals for behavior control will continue to be controversial.

Sometimes, aversive techniques are considered for children experiencing learning and behavior problems. For example, such procedures as punishment, implosion, or flooding, might be used. Increasingly, techniques involving aversive procedures have come under judicial review. Typically, the courts have become involved when individuals might be exposed to "cruel and unusual" punishment. In some cases isolation procedures have been ruled illegal (*New York State Association for Retarded Children v. Rockefeller*). When aversive procedures, such as time out, are employed, clients must have access to food, lighting, and personal hygiene facilities (*Hancock v. Avery*). Moreover, the courts have also ruled that isolation may only be used for behaviors leading to physical harm and/or destruction of property (*Morales v. Turman; Wyatt v. Stickney*). Even then, these procedures must be monitored by professional staff. More severe punishment procedures such as electric shock can only be used in those unusual circumstances where the client is engaging in severe self-destructive behavior (*Wyatt v. Stickney*).

Generally, there have not been court rulings on many of the specific aversive procedures that might be used in the treatment of children. An important consideration in treatment of learning and behavior problems is that aversive procedures may in most cases, be inappropriate. Thus, the practitioner should consider the range of possible nonaversive procedures that could be used in therapy.

Personal Rights

A major issue related to the control of behavior is personal or human rights in therapy (Schwitzgebel & Schwitzgebel, 1980). Most interventions involve a definite control of behavior and so issues of personal rights and freedom are raised. This is especially an important issue with children because therapists typically intervene for the good of the child. But, as Ross (1980) has questioned, can the professional always be trusted to protect the child's rights and best interests? The issue of the child's rights is somewhat different than that of an adult (Ross, 1980). First of all, the child is usually brought to treatment or referred to treatment by an adult. Typically the child does not volunteer for therapy. Second, the child usually does not decide what the goals of treatment should be, or what treatment should be, or when treatment should terminate. Issues like these have prompted some individuals to recommend a "Bill of Rights" for the child/client (Ross, 1974, 1980; Koocher, 1976). The rights include four basic principles:

The Right to be Told the Truth. The basic premise of this principle is that the child should not be deceived. The child should be told the truth regarding the purpose for treatment and what it will involve.

The Right to be Taken Seriously. The child's perspective in his/her problem and his/her option on various issues should be seriously considered and not dismissed because a child is speaking.

The Right to Participate in Decision Making. The child should be included in decisions that are made regarding his/her treatment program. Adopting such a strategy does not imply that the child's perspective will determine the final decision. Yet, the child's perspective should be considered in with others involved in the therapeutic program (e.g., parent, teachers).

One method to protect the rights of the child would be to develop a contractual arrangement for services (e.g., Kazdin, 1980; Schwitzgebel, 1975; Schwitzgebel & Schwitzgebel, 1980; Stuart, 1977). Some advantages of a contractual arrangement include the following:

1. The contract spells out mutual goals and committments.
2. Contracts can be used in a variety of settings.
3. Contracts encourage negotiation of priviledges and responsibilities.
4. Contracts reduce disagreements over what is to take place in therapy.

Selection of Intervention

In intervention with children's learning and behavior problems, the professional must consider the relative efficacy and the efficiency with which the problem can be solved (Wilson & O'Leary, 1980). Prior to implementation of any intervention program the professional should consider several issues (Morris & Brown, 1982).

Is the treatment program consistent with the available treatment literature? (If it contains any novel intervention approaches and/or if a new treatment method is being proposed where there are no data to support its efficacy, the therapist may want to propose the treatment as an experimental procedure).

Is the program consistent with the overall treatment objectives for the child and is it in the child's best interests?

Does the program involve the least restrictive alternative program for the child?

Can the program be carried out easily given the number of staff available and the level of staff training and competence?

Will the child's progress be monitored using a specific procedure and will the child be observed closely for possibly adverse side effects of the program?

Have the staff been trained to a criterion level to ensure the provision of quality treatment?

Has informed consent been obtained from the client and/or the parents/guardians?

Each of these issues pose special problems for the professional involved in treating children. Each will be discussed as it relates to ethical and legal issues in the field.

Available Treatment Literature. In our analysis of the ethical factors governing the selection of treatment procedures for children, the choice of one treatment over another should be based on a careful review of the literature (McNamara, 1978). To help the therapist decide on what type of treatment to employ, the following series of questions can be helpful:

1. How effective is a given technique for the presenting problem?

2. How costly is the procedure relative to other techniques known to be of equal benefit?
3. Are there any negative side effects associated with the procedure?
4. How durable are the effects of the treatment?
5. Does the treatment have a high probability of being implemented by the therapist, client, and/or provider?

Some of these questions are obviously research questions and relate to methodological and conceptual work. It is clear that the professional should examine the research literature to answer many of the questions that will arise in this area of ethical concern.

Intervention Objectives. A primary intervention objective is to change the behavior (eliminate the problem) so that professional involvement can be terminated. Generally, intervention goals should be individual and specific to the problem of concern. Several questions can help guide the professional toward more effective intervention objectives (Martin, 1975, pp. 69-70):

1. Does your program have a concrete, objectively stated goal?
2. Is it directly related to the reason the individual was brought to your attention?
3. When it is achieved, can your involvement with the client be terminated?
4. Will the change benefit the individual more than the institution?
5. Can the goal be achieved?
6. Is the goal a positive behavior change rather than a negative behavior suppression?

7. Does the goal involve changing a behavior that is actually constitutionally permissible?

A question that can be raised in intervention programs with children is to what extent can the child participate in objectives and goals. This question essentially raises the issue of the competence of the child to make important decisions bearing on interventions. Some information related to the child's ability to consent to interventions has come from Grisso and Vierling (1978). These authors reviewed the developmental research literature and reached the following conclusions:

1. There may be no circumstances that would justify sanctioning independent consent by minors under 11 years of age, given the developmental psycholoigcal evidence for their diminished psychological capacities.
2. There appear to be no psychological grounds for maintaining the general assumption that minors at age 15 and above cannot provide competency consent.
3. Ages 11-14 appear to be a transition period in the development of important cognitive abilities and perceptions of social expectations, but there may be some circumstances that would justify the sanction of independent consent by these minors for limited purposes, especially when competence can be demonstrated in individual cases (p. 424).

Unfortunately, these conclusions were not based on research directly bearing on intervention decisions in real life situations (Melton, 1981). With the lack of such a data base, it is likely that the courts would accept the somewhat arbitrary age of majority of 16 years for informed consent. Melton (1981) noted that youngsters are usually competent to give consent at least after age 15. Yet, this would depend on individual differences, cognitive abilities, and the unique circumstances of the problem. In any case, the professional should determine the actual capacity to give consent and plan therapeutic goals.

Least Restrictive Alternatives. The least restrictive alternative applies to both consideration of alternatives to committment and alternatives for interventions available (Schwitzgebel & Schwitzgebel, 1980). For example, in the Wyatt case it was noted that individualized treatment plans are necessary for an effective program "and each plan must contain a statement of the least restrictive treatment conditions necessary to achieve the purposes of commitment." A major goal in providing services for children experiencing learning and behavior problems should be to select an intervention that is relatively nonintrusive or restrictive. Providing the child with the least restrictive alternative intervention will promote the opportunity to change under minimally intrusive and restrictive conditions. The terms "restrictiveness" and "intrusiveness" refer to "methods that involve a high degree of obvious external control, especially those based on aversive control" (p. 289). Friedman (1975) has defined "restrictiveness" in terms of "a lost of liberty" and "intrusiveness " in terms of placing a person at risk, using force to modify the behavior of a person, invading someone's body, or the loss of personal autonomy.

In addition to definitions of intrusiveness or restrictive intervention methods, two sets of criteria have been proposed; one of these is for the instrusive nature of a particular intervention (Shapiro, 1974), and the other has been developed to evaluate the intrusiveness of behavioral and other procedures with prisoners and psychiatric patients

(Speece, 1972). Shapiro (1974) proposes the following six criteria:

1. Is the effect of the therapy procedure reversible?
2. Does the effect of therapy result in behaviors which are judged to be maladaptive and/or inconsistent with "normal" functioning?
3. How quickly does the behavioral change occur following the initiation of the therapeutic procedure?
4. To what extent can a person avoid behaving in the planned manner?
5. What is the duration of the resulting behavior change?

The criteria proposed by Speece (1972) also include components that can be applied to interventions in educational settings:

1. The nature and intensity of the collateral behaviors and other side effects which develop as a result of the procedure, as well as the duration of the effect on the targeted behavior.
2. The extent to which an uncooperative client can avoid the procedure, i.e., exert countercontrol vis-a-vis the therapeutic procedure;
3. The extent to which the procedure involves the introduction of physical contact with the body of the client.

It seems clear that procedures advocated in the literature on treatment of children are sometimes intrusive and restrictive.

Generally, the principles associated with the concept of least intrusive or restrictive intervention necessitates that more intrusive methods be applied only after less intrusive methods have been demonstrated to be ineffective. Morris and Brown (1982) have proposed a system based on the provision of services to the mentally retarded, but which is useful in the treatment of children experiencing other learning and behavior problems. This system, described in Table 8.1, varies along both the dimensions of restrictiveness/intrusiveness and aversiveness (as defined in terms of the frequency, intensity, duration, and topography of the aversive intervention introduced to decrease the child's target behavior). In this system, professionals should demonstrate that Level I interventions have been ineffective in controlling a behavior before proceeding to implement Level II treatments. In a similar manner, prior to implementation of Level III procedures, the professional would have to demonstrate that Level II procedures were ineffective. These considerations must also be employed within the context of other ethical imperatives (e.g., human rights, informed consent).

Available Professionals and Training. An important issue in treating children involves the consideration of who will carry out the program and whether those individuals are trained (qualified) to do so. Even though a specific procedure might be available for use in treatment, individual(s) qualified in its delivery must be available for either the direct service or supervision of the individuals who will carry out the program. Many of the procedures used with children might appear deceptively simple, but in reality are quite complex when correctly implemented. For example, Agras

Table 8.1

Proposed Levels of Restrictiveness/Intrusiveness and Aversiveness of Behavior Modification Procedures with Mentally Retarded Persons

Level I Procedures

Reinforcement
Shaping
Modeling
Token Economy System
Ecological/Behavioral Engineering
Self-Control
Reinforcement of Incompatible Behaviors
Extinction

Level II Procedures

Contingent Observation
Exclusion Time-Out
Response Cost
Contact Desensitization

Level III Procedures

Overcorrection
Seclusion Time-Out
Negative Practice
Satiation
Physical Punishment

Source: Morris, R. J., & Brown, D. K. Legal and ethical issues in behavior modification with mentally retarded persons. In J. Matson and F. Andrasik (Eds.) *Treatment issues and innovations in mental retardation.* New York: Plenum Publishing Co., 1982. Reproduced by permission.

(1973) in discussing the qualifications of a well-trained behavior therapist noted that this individual:

> ...must have knowledge of the principles underlying behavior modification, experience in the application of such knowledge to human behavior problems, and experience in the experimental analysis of deviant behavior, both for research purposes and as an approach to the on-going evaluation of clinical care. (S)he must also, however, demonstrate certain less well-defined characteristics, usually referred to as general clinical skills (p. 169).

As noted by Wilson and O'Leary (1980) such "clinical skills" are typically acquired through formal graduate training. Professional organizations have, in some cases, developed guidelines for the delivery of services [(see Chapter 9, e.g., psychologists would follow the Speciality Guidelines for the Delivery of Services (APA, 1981) in the areas of psychological speciality]. For the APA these include the guidelines in the area of clinical, counseling, industrial/organizational, and school). In and of itself, graduate training will certainly not guarantee competence. Individuals may also be certified or licensed by state boards. Moreover, individuals may also belong to professional organizations that provide certain status or recognition for competence in a certain area (e.g., diplomate status in APA).

But, there are issues of quality intervention that extend beyond the professional's skills. Even if the professional is well-qualified to deliver services, many programs for children's problems are implemented by paraprofessionals and/or the child's providers (e.g., parents). In such cases, the professional will be involved in supervision of those individuals providing the intervention services. At least two issues are important here, namely, training and monitoring of the individuals (cf. Martin, 1975). In some cases, such as in institutional programs, individuals might be selected for intervention implementation. In other cases, the professional will need to ensure that these individuals are trained. For example, in some institutions this training might take the form of orientation, pre-service training, implemented in-service training, and planned in-service training (see Martin, 1975, pp. 110-112). Certainly, these procedures appear necessary and desirable in most institutional settings. Yet, implementation of these strategies in some schools and especially in home settings may prove especially difficult. Nevertheless, some formalized attempt must be made to provide the individuals impelementing the program some sort of training to carry out the task.

In addition to a training component, supervision of the ongoing services will be necessary. Such supervision is aimed at ensuring that the program is being implemented as intended and to revise it given that it is not working. In such cases, data must be gathered on the client (see below).

Program Monitoring. In order to ensure that the intervention is being implemented correctly and that it is having desirable effects it must be monitored by the professional or his/her designers. Some writers believe that ongoing evaluation of clinical services is essential (Barlow, 1980, 1981). Methods through which this can be accomplished will vary from case to case, but would include self-report inventories and checklists, self-monitoring, direct observation by parents, or teachers (e.g., Nelson, 1981). For many practitioners this data gathering operation will prove especially difficult and costly. Yet, some attempts must be made to gather data on client outcome.

Aside from monitoring data to determine if the intervention is having an impact, it is desirable to gather data to monitor any side effects, both positive and negative. In the case of positive side effects, an especially effective intervention will usually result in positive behaviors for the child. The child may return to the regular classroom, improve academic performance, develop new friends, acquire new social skills, and so forth.

Monitoring undesirable side effects is also important with certain intervention procedures. For example, implementation of certain aversive procedures such as implosive therapy may result in development of undesirable behavior, such as avoidance. In addition to monitoring certain undesirable behaviors of the child, the professional should consider the potential negative influences of a program on the parents and/or siblings. Moreover, a desirable change in the child's behavior could result in a negative change in a parent's or other sibling's behavior.

Gathering data to monitor the intervention program and its side effects should not necessarily be regarded as research, for these two activities are different on both methodological and conceptual grounds (cf. Kratochwill & Piersel, in press). Yet, by gathering some data on the client, the professional becomes accountable to the consumer (Wilson & O'Leary, 1980). Thus, rather than reported "success" through subjective means, the professional may be able to provide some type of credible data to document change.

Informed Consent. Some of the issues involved in informed consent have already been elucidated in the context of assessment. Similar concerns must be addressed in implementation of an intervention program.

Three major issues can be raised regarding informed consent in therapeutic work with children, namely competency to give consent, freedom from constraint, and clarity in the information given (Stuart, 1981).

Competency to give consent. As noted above, the issue of when and if children are competent is characterized by considerable controversy. Usually, parents or guardians must play a primary role here, since the child might be judged as incompetent to give consent. Yet, in other cases, especially where guardianship has not been established, a group or committee external to the institution or circumstances should assign an advocate to the child to assist him/her in determining whether the intervention program is acceptable (Morris & Brown, 1982; Ross, 1980).

Even in the case of adults, the issue of when an individual is competent to give consent is quite subjective. Some authors have characterized competency as the appearance that the individuals know what he/she is doing (e.g., Hardisty, 1973) or if he/she seems to know what they are doing in a layman's sense (e.g., Roth, Meisel, & Lidz, 1977). In some cases the courts have upheld that persons are considered legally competent unless it can be proven otherwise (Lotman v. Security Mutual Life Insurance Co., 1973).

Five different methods have been proposed for determining competence (Roth, Meisel, & Lidz, 1977, reviewed by Stuart, 1981, pp. 719-720):

1. A person may be judged competent if he/she shows a clear desire to participate in the activity.
2. Competence can be inferred from the judgment that the person has made a "reasonable" choice (Friedman, 1975).

3. Competence can also be inferred from the belief that participation in a program is based on a rational process (Stone, 1975).
4. Competence is inferred when the person displays the ability to understand the nature of the intervention.
5. The competence of the individual is evaluated by assessing the actual level of understanding of the procedure.

Procedures for selecting one of these tests of competence have been proposed (Roth, Meisel, & Lidz, 1977). The test ranges from the least stringent (consent through participation) to the most stringent (demonstrating understanding). It would appear that such "tests" could be applied with children. Nevertheless, these criteria would need to withstand the scrutiny by the courts; currently they appear quite subjective (see Stuart, 1981 for further discussion of these issues).

Freedom from Constraint. Coercion occurs ..."when false or incomplete information is given about proposed procedures, when nonparticipation is punished in a way other than by simple loss of the potential benefits of participation, or when compliance is obtained through physical coercion" (Stuart, 1981, p. 721). When these factors enter into the intervention process truly informed consent is not possible. Constraint is especially worrisome in the case of children.

Unfortunately, even the alternative of assigning an advocate to the child does not allow him/her to refuse intervention. The issue of intervention refusal is an important one especially with children and specifically in the case of a severe behavior problem. It would appear that children should have the right to refuse intervention, but the right to provide intervention also exists. When a compromise cannot be developed, formal legal rulings may be the only alternative (Morris & Brown, 1982).

Some of the more salient legal problems with refusal of intervention include the following (see Stone, 1975):

1. The client's competency to decide whether or not to refuse treatment.
2. Procedures for obtaining informed consent of a severely disturbed but legally competent individual.
3. Handling objections on religious grounds.
4. The civil libaility of a practitioner if a client who has refused treatment injures him/herself or others.
5. Increased cost to taxpayers of individuals who refuse less expensive treatment and insist on more expensive ones (Schwitzgebel & Schwitzgebel, 1980 p. 53).

Generally, the individual involved in treating children's learning and behavior problems must determine the level of coercion in each case and minimize it within the professional relationship.

Clarity of the Information Given. The clarity of the information given can influence the degree to which consent is truly informed. Generally, information should be complete and communicated in a clear fashion. For therapeutic purposes, a multiple-pact consent form can be employed (Martin, 1975; Miller & Willmer, 1974).

Issues in Research

Experimentation with children experiencing learning and behavior problems also raises a number of ethical and legal considerations. Many of the issues that are raised in research are similar to those that have been presented within the context of assessment and intervention. Yet, some special issues emerge in research simply because research is the primary activity. Several considerations have been advanced (Kazdin, 1980). First, since experimental research usually requires manipulations of variables, subjects could be exposed to certain conditions that are harmful or stressful to them. For example, a child experiencing a severe emotional problem might be exposed to an intervention that causes a great deal of stress and anxiety. Whether or not a child should be exposed to interventions that cause stress raises both ethical and legal issues.

Another consideration is that in research information may be withheld from the child. Providing information may reduce the efficacy of the intervention or conflict with the goals of experimentation. Yet, withholding information from the child and or the parents may not meet informed consent guidelines.

Third, the actual data collection that occurs in the typical research process may involve the privacy of the child and his/her parents. For example, as part of data monitoring in a home setting, certain private and personal information might be revealed.

Fourth, some of the methodological requirements of research may conflict with intervention objectives of the child. For example, in between-group research some subjects might be assigned to a condition that provides no intervention or to a condition in which the child receives an intervention known in advance to be less effective than another available method.

Fifth, the differential status between the investigator and the child raises ethical concerns in that the child becomes vulnerable to possible abuses. Children might not object to some intervention that an adult would readily object to. Because children are frequently in a "non-power" position, they are more likely to suffer certain types of abuses.

There has been growing recognition that the protection of human subjects in research is necessary. A number of laws regulating research with human subjects have been proposed. The Nuremberg Code (1946), Declaration of Helsinki (1964) and the Institutional Guide to DHEW Policy on Protection of Human Stubjects (1975) all testify to the recognition that human rights are important in research. Some of these are specific to children. In addition to these, the APA (1981) under Principle 9 of the Ethical Principles of Psychologists has provided 10 guidelines for research with human subjects.

In this chapter we review some major legal and ethical issues that have been raised in the conduct of research with human subjects. Our discussion is not comprehensive, but is designed to elucidate some specific issues that emerge in assessment and intervention research on children. For a more detailed discussion of ethical and legal issues in research, the reader should consult several sources (e.g., Bersoff, 1978; 1979; Brady, 1979; Kazdin, 1980; Kelman, 1971; McNamara & Woods, 1977; Schwitzgebel & Schwitzgebel, 1980).

Informed Consent

The informed consent doctrine first emerged as a formal rule for the physician-patient (Bersoff, 1978). Essentially, the issues raised in assessment and intervention apply in research. Yet, by virtue of labeling one's activity as "research", some special concerns emerge. For example, typically the investigator must make a formal request for conducting the research and have a research proposal reviewed by an independent committee. Some special problems that may arise in this area involve the capability of children to give consent (see above discussion). Also providing advanced knowledge about a particular intervention may be difficult if it is a new technique and/or little is known about its influence from previous research. Such factors may make a truly knowledgeable decision impossible (Kazdin, 1980). A number of suggestions have been advanced for demonstrating that potential research subjects are informed prior to providing consent to participate in a research program. For example, the use of a two-stage (Miller & Willner, 1974) and a three stage (Stuart, 1978, 1981) consent form have been proposed. Grabowski, O'Brien, and Mintz (1979) proposed a system based on well-constructed information forms and correlated multiple choice items. The materials include a description of the consent procedures, a statement of purpose, description of experimental procedures and alternatives, and statements stipulating that withdrawal is an ongoing option. Despite these options, providing such information has been shown to influence both the subject's willingness to participate (Stuart, 1978) and the potential results (Grunder, 1978).

When precautions have been taken to inform the subject, questions have also been raised over the meaningfulness of the activity (Palmer & Wohl, 1972). Subjects may forget that they signed a consent form or indicate that they did not understand the purpose of the study.

Sometimes researchers may not inform subjects that they will be randomly assigned to conditions assuming that they will then refuse to participate. McLean (1980) studied the effects of informing clinically depressed subjects that their treatment assignments were made on a random basis, in terms of their willingness to consent to intervention. He found that none of the 104 subjects who were informed of random assignment refused to participate in the program. Also, there was only a negligible effect in subjects' willingness to consent between the informed and uninformed conditions. McLean (1980) noted that the issue of random assignment may be less critical than other issues raised over informed consent procedures.

Deception

Closely related to the informed consent notion is the use of deception. Technically, the true informed consent might be held to free of any deception. Yet, the issue in research is whether or not the deception is justified in light of the benefits that might ensue from the research (Kazdin, 1980). Although the scientific contributions of a study may determine if the deception can be justified (Kelman, 1968), the benefits of a particular study can be difficult to assess, especially when the researcher has vested interests in the investigation (Kazdin, 1980). Based on issues such as these, the justification for deception in research depends on several considerations:

> First, the scientific investigation must merit the type of deception that is used. Whether or not the deception is merited is, however, a subjective judgment that requires reliance on persons other than the possibly biased investigator. Second, there must be assurances that alternative methods of investigation that would produce the information proposed in an experiment that uses deception is entirely an empirical matter. Researchers may argue in all honesty about the extent to which deception is essential. Third, the aversiveness of the deception itself bears strongly on the justification of the study. Deceptions vary markedly in degree, although ethical discussions usually focus on cases where subjects are grossly misled about their own abilities or personal characteristics. Finally, the

potential for and magnitude of the harmful effects of the deception on the subjects also dictate whether the deception would be justified. Whether an experiment using deception is justified needs to be weighed carefully. Increasingly, research that seriously misleads the subject simply is not permitted (Kazdin, 1980, p. 390).

Debriefing

Once any deception has been employed in research, it is the responsibility of the professional to describe the nature of the experiment, that is, to debrief the subject regarding the purposes of the study and what was done in the study. A major purpose for this debriefing activity is to minimize any stress or problems that may have been a function of the actual deception(Kazdin, 1980; Kelman, 1968).

Although debriefing appears to be an important activity for the researcher, many unanswered questions are raised regarding this particular tactic. For example, it is possible that the debriefing activity does not resolve the problems that were raised for the client. In this regard, it is possible that a youngster who is exposed to an aversive situation in a study when debriefed may not, in fact, feel more comfortable. It is quite possible that the subject may feel hostile or fearful toward the experimenter no matter how much debriefing takes place.

Questions might also be raised as to when the debriefing should take place. For example, it might occur immediately after the subject participates or after all subjects have participated in the experiment. In the latter case it might be assumed that debriefing all children at the same time at the end of the study would minimize communication among subjects if this is an issue. However, this would need to be weighed against the potential negative effects of having the child experience a period of time under which the deception was employed. Another issue is that even if attempts are made to debrief the subject regarding the nature of the study, it cannot always be assumed that the person uunderstands what was done and why it was done. In some respects, the same problems that occur in debriefing are those that emerge in the informed consent issue. Particularly with young children the issue of understanding the debriefing activity might be raised. As Kazdin (1980) has noted, the investigator employing any deception must demonstrate that debriefing activities were in fact successful. It behooves the investigator to make sure that the debriefing activities are systematic, well controlled, and monitored.

Summary and Conclusions

In this chapter, we have provided an overview of the ethical and legal considerations in assessment, treatment, and research of minority and nonminority children. The issues raised with respect to children apply to research, intervention, and assessment in these areas and also extend beyond work in this area whenever children are involved. The work with children in assessment, intervention and research involves considerations of several factors including law, ethics, and morality. These influences provide a conceptual guide for the professional involved in work with children's learnning and behavior problems. As we noted, laws have provided one of the strongest influences on professional behavior, but in many cases laws have yet to be enacted for specific situations and, in many cases, they have been postscriptive rather than prescriptive. As a second source of influence, ethics have usually been developed as guidelines for individuals working in the field. In practice, rights and ethics overlap and it is important for the clinician to consider the various ethical guidelines for professional behavior across various disciplines to which he/she adheres. Finally, moral principles have provided guidelines for conduct that transcends specific laws and ethical codes. These typically refer to absolute assumptions about the rights and responsibilities of individuals.

In assessment work with children experiencing learning and behavior problems, a number of issues have emerged. Specifically, some criticisms of assessment including invasion of privacy, crating an unfavorable atmoshpere, developing labels, engaging in discriminatory practices, have all been advanced. Each of these considerations must be noted in any assessment work. Several influences from the legistlative and judicial areas, as well as professional associations have been raised for guiding assessment activities. These were reviewed as they appeared relevant for assessment of children's learning and behavior problems. The Larry P. and PASE decisions were examined closely. The different conclusions reached by the judge in each case was suggested to be a result of the definition of bias each judge adopted.

A number of issues have been raised in intervention efforts for children. In this area of learning and behavioral disturbance, issues have been raised in the control of behavior, personal rights, and selection of interventions. When selecting a particular intervention strategy, the professional must consider the available literature, specific intervention objectives, least restrictive alternatives, available professional staff and training of these individuals, monitoring of the program established, and informed consent. The informed consent notion strongly advocated in any intervention efforts is a complex one and not easy to address. Issues that were reviewed here included the competency to give consent, freedom from constraint, and clarity of the information given to the child and/or his/her providers.

Several issues involved in intervention research on children's learning and behavior problems, were also reviewed. These included informed consent, invasion of privacy, deception and debriefing. Each of these issues was reviewed in the context of some issues from the professional literature reviewed in earlier sections of the report. It is hoped that future assessment, intervention, and research activities of individuals working with children experiencing learning and behavior problems will be guided through considerations raised in this chapter.

Chapter 9

The Influence of Professional Organizations

Professional organizations have influenced assessment activities in general and issues in the area of test bias specifically. The impact has been in several areas (Oakland & Laosa, 1977). First of all, various professional organizations have become involved in making public statements on testing and test bias. Indeed, as will be emphasized below, some groups have taken a formal position against standardized tests in educational settings. Second, certain professional groups have published guidelines to accompany various assessment practices. Such guidelines often specify the nature of professional conduct in the choice, administration, and use of tests and assessment practices. Third, some professional groups have been involved in certifying and licensing individuals who offer these psychological or educational assessment services.

In this chapter we review some of the professional groups that have been active in establishing positions and/or have prepared documents related to assessment practices. It should be emphasized that although a number of professional groups/organizations have considered issues relevant to bias in assessment, only a few have provided any formal guidelines related to practice. The professional organizations that have provided standards for assessment/intervention are listed in Table 9.1. This list is by no means exhaustive but should alert readers to consider the existing standards and guidelines from their own professional organizations. A review of other organizations and societies that have developed ethical policies can be found in the AAAS Professional Ethics Project (1980). This document provides a review of guidelines developed in both the physical and the social sciences.

Groups Representing "Marxist": Opposition to Tests

Several self-identified "Marxist" groups have come out in opposition to various psychological tests. Jensen (1980) reviewed some perspectives in this area so we will only present a brief overview. As Jensen (1980) notes, there is nothing intrinsic in original Marxian theory that would be in opposition to mental tests. Moreover, although IQ tests were once disdained in the Soviet Union, testing is still apparently common in this country. Nevertheless, some opposition to tests, particularly in the study of individual differences, have been advanced (Teplov & Nebylitsyn, 1969). Presumably, Soviet psychologists have relied somewhat less on tests than those psychologists in the United States.

Marxists outside the USSR have expressed objections to ability tests (e.g., Lawler, 1978, Simon, 1971). One

Table 9.1

Professional Organizations Who have Provided Standards/Guidelines for Assessment/Practice

American Educational Research Association (AERA)

American Personnel and Guidance Association (APGA)

American Psychological Association (APA)

Association of Black Psychologists (ABP)

Association for Advancement of Behavior Therapy (AABT)

National Association for the Advancement of Colored People (NAACP)

National Association of School Psychologists (NASP)

National Council of Measurement in Education (NCME)

National Education Association (NEA)

Society for the Study of Social Issues* (SSSA) perspective on this is presented by Simon (1971):

---

*The SSSA is Division 9 of the American Psychological Association.

(Source: Kratochwill, T. R., Alper, S., & Cancelli, A. A. Nondiscriminatory assessment: Perspectives in psychology and special education. In L. Mann & D. A. Sabatino (Eds.), The fourth review of special education. New York: Grune & Stratton, 1980. Reproduced by permission)

> Since, in a class society, on average, the higher the social status, the greater the likelihood that test questions of the kind described can be answered; a test standardized this way is bound to set standards of "intelligence" which are largely class differences disguised. It is an inescapable fact that the middle class child will always tend to do better than the working class child, as a necessary result of the way in which the tests are constructed, validated, and standardized (p. 78).

Groups Representing Minorities

Several different professional organizations that represent minority groups in the United States have made formal statements regarding the use of standardized tests. The Association of Psychologists for La Raza (APLR), an organization for Chicano psychologists, does not have an official position on minority assessment. However, the president of the association responded to the APA report on "Educational Use of Tests with Disadvantaged Students" (Cleary et al., 1975). Although the report stressed fair assessment practices, Bernal (1975) pointed to various oversights in the report:

> The key arguments of many critics of extant testing and test development procedures have not been discussed or answered, recommendations for

> improving test development with and for minorities have not been set forth. The blame for bad testing appears to have been shifted to the practitioner, and the schools seem to be the only institutional villains in the story. In short, the classic "Type III" errors were made by a committee that lacked minority membership to articulate minority perspectives: Not enough of the right research questions and issues of interest were raised. As a result, the document generally has become an apologia for testing (p. 92).

Professional groups representing black minorities have been somewhat more active in their opposition to certain testing practices. For example, the National Association for the Advancement of Colored People (NAACP) held a conference on minority testing in 1976. The report (Gallagher, 1976) pointed to uses and misuses of tests, psychometric issues, public policy, and a code to help ensure the fair use of tests.

Perhaps the most influential group in the testing arena has been the Association of Black Psychologists (ABP) with their call for an immediate moratorium on the use of psychological tests with children from disadvantaged backgrounds. In a subsequent report, Williams (1971, p. 67) noted that ability tests:

1. Label black children as uneducable,
2. Place black children in special classes,

3. Potentiate inferior education,
4. Assign black children to lower education tracts than whites,
5. Deny black children higher educational opportunities, and
6. Destroy positive intellectual growth and development of black children.

The ABP has continued to be quite active in their opposition to standardized tests used for special education classification in schools. The organization has spear-headed suits against school districts (see discussion in Chapter 8). Also, in part, as a response to the ABP's proposed moratorium on the use of psychological tests with blacks, the American Psychological Association's Board of Scientific Affairs formed an ad hoc committee to investigate the validity of testing in educational settings. The committee report covered a broad spectrum of issues, including theory of human abilities, test misuse and misinterpretations, evaluation of the "fairness" of tests in use, and alternatives to commonly used intellectual tests (Cleary et al., 1975).

Reaction to the report from the ABP was quite negative. Jackson (1975) noted:

> In this writer's judgment the report is blatantly racist. It continues to promulgate the notion of an "intellectual deficit" among black people, seeks to treat all disadvantaged in a similar manner,and employs a definition of "fairness" which is

> intrinsically unfair. It attempts to describe the retesting functions in a seemingly educationally desirable manner when in fact these functions serve to sustain and maintain the status quo while systematically prohibiting black self-actualization and self-determination and promulgating exclusion of blacks from the American mainstream. The committee appears to have ignored the wealth of work of black psychologists in this area. To discuss and scientifically discount is one thing; to totally ignore is racism at its arrogant worst (p. 88).

The APA committee chairman (Humphreys, 1975) wrote a rebuttal to the ABP's reaction. Humphreys (1975) noted:

> The authors of the report also believe that test scores properly interpreted are useful. We do not and cannot support a moratorium on testing in the schools. Furthermore, many useful interpretations of test scores can be made without appreciable loss of accuracy in the absence of information about race, ethnic origin, or social class of the examinee. Whether demographic membership is needed is an empirical matter and not one decided on the basis of ideology (p. 95).

Nevertheless, the ABP has rejected the APA report and noted that a moratorium is no longer enough; what is needed is government intervention and sanctions against testing practices.

American Personnel and Guidance Association

Guidance counselors and associated personnel are sometimes involved in testing practices. This is often associated with vocational or career assessment and may involve minorities. At the 1970 annual convention of the American Personnel and Guidance Association (APGA), the Senate adopted a resolution in which concern was expressed over minority group testing. Thereafter, the Association for Measurement and Evaluation in Guidance (AMER, a division of APGA) prepared a position statement on the use of tests, and with the assistance of AMEG, APGA, and the National Council of Measurement in Education (NCME), a paper was adopted as an official position of those organizations (AMEG, 1972). In the document it was noted that:

> Professional associations, including the measurement societies, do not have the authority to control intentional discrimination against particular groups, though individual members acting in accordance with their own consciences may bring to bear such powers as their positions afford them (AMEG, 1972, p. 386).

In the document it is also stated that issues relating to test misuse should go through the court system, boards of education, civic service commissions, and other public groups.

National Education Association

The National Education Association (NEA) has come out against standardized tests. In 1972, the NEA's Center for Human Relations held a three-day national conference in Washington, D.C. The theme of the conference was "Tests and Use of Tests--Violations of Human and Civil Rights (Bosma, 1973). Individuals attending the conference were asked to complete a questionnaire, including such items as:

> IQ tests are not perfectly accurate nor are they a perfect indication of potential.
>
> The IQ test is a measure of experience and learning rather than a measure of inborn ability.
>
> Most standardized tests are tests of developed abilities rather than measures of potential.
>
> Given the possible negative effects of standardized tests, which of the following actions do you believe should be taken?
>
> (a) Eliminate the use of standardized tests entirely.
>
> (b) Intensify efforts to develop culture free tests.
>
> (c) Curtail the use of standardized tests except for research purposes.
>
> (d) Conduct an intensive educational program to prevent misuses of tests (Cited in Jensen, 1980, p. 13).

Following the meeting the NEA policy-making

Representative Assembly passed three resolutions (Oakland & Laosa, 1977, pp. 22-23):

1. To encourage the elimination of group-standardized intelligence, aptitude, and achievement tests until completion of a critical appraisal, review, and revision of current testing programs;

2. To direct the NEA to call immediately a national moratorium on standardized testing and set up a task force on standardized testing to research the topic and make its findings available to the 1975 Representative Assembly for further action; and

3. To request the NEA task force on testing to report its findings and proposals at the 1973 Representative Assembly.

In 1973 the NEA task force again called for a national moratorium on standardized testing until 1975. The NEA Representative Assembly also reviewed the moratorium resolution on testing, suggesting that tests should not be used in a manner that denys students full access to equal educational opportunity.

National Association of School Psychologists

School psychologists are nearly always involved in assessment of children in educational settings. Many of the children who are referred for psychological or special educational services represent various minority groups. The National Association of School Psychologists (NASP) is one professional organization representing practicing and academic school psychologists in the United States and some foreign countries. Because school psychologists are frequently in an extremely sensitive position in school assessment practices, the NASP delegate assembly (NASP, 1978) has adopted a number of resolutions that have a bearing on assessment (e.g., Resolutions 3, 6, and 8). For example, Resolution 3 notes that school psychologists should protect children, especially those in minority groups, from abuses through the malpractice of school psychology.

Resolution 6 is more explicit in expressing the position that blacks and other minority groups do not manifest an inferiority in intellectual functioning based on so-called genetic characteristics. The NASP has argued that there is inadequate scientific support for genetic differences in intelligence among groups and that research into the issue is needed.

Resolution 8 notes that:

> Individuals of different socio-cultural backgrounds differ in their readiness to succeed in school; that professional members of minority groups have indicated that it is a disservice to minority individuals to suggest that they need not do well on tests or achieve a basic education; and that objective measures are less biased than subjective judgments in assigning children to special programs in schools (p. 104).

In addition to these resolutions, some specific suggestions for standards relating to professional involvement, assessment standards, standards for parent and/or student involvement, standards for educational programming and follow-through, and training standards follow these resolutions (NASP, 1978, pp. 105-107).

Association for Advancement of Behavior Therapy

The Association for Advancement of Behavior Therapy (AABT) represents practice and research interests of behavior therapists. In May 1977, the Board of Directors of the AABT adopted "Ethical Issues for Human Services." The guidelines do not mention issues related to test bias. In fact, the statements in the guidelines are conceptualized within the domain of treatment (see Table 9.2).

Within contemporary behavior therapy, assessment and treatment are conceptually linked (cf. Kratochwill, 1980, 1982) and so it is possible to apply any one of the guidelines within the context of assessment practices. Nevertheless, the AABT has shown increasing interest in assessment practices, as reflected in the formation of the journal Behavioral Assessment. Whether or not the organization will make any formal statements on test bias remains to be seen.

The AABT guidelines take on special significance in

Table 9.2

Ethical Issues for Human Services

The questions related to each issue have deliberately been cast in a general manner that applies to all types of interventions, and not solely or specifically to the practice of behavior therapy. Issues directed specifically to behavior therapists might imply erroneously that behavior therapy was in some way more in need of ethical concern than non-behaviorally-oriented therapies.

In the list of issues, the term "client" is used to describe the person whose behavior is to be changed, "therapist" is used to describe the professional in charge of the intervention; "treatment" and "problem," although used in the singular, refer to any and all treatments and problems being formulated with this checklist. The issues are formulated so as to be relevant across as many settings and populations as possible. Thus, they need to be qualified when someone other than the person whose behavior is to be changed is paying the therapist, or when that person's competence or voluntary nature of that person's consent is questioned. For example, if the therapist has found that the client does not understand the goals or methods being considered, the therapist should substitute the client's guardian or other responsible person for "client," when reviewing the issues below.

A. Have the goals of treatment been adequately considered?

Table 9.2 (Cont.)

1. To insure that the goals are explicit, are they written?
2. Has the client's understanding of the goals been assured by having the client restate them orally or in writing?
3. Have the therapist and client agreed on the goals of therapy?
4. Will serving the client's interests be contrary to the interests of other persons?
5. Will serving the client's immediate interests be contrary to the client's long-term interest?

B. Has the choice of treatment methods been adequently considered?

1. Does the published literature show the procedure to be the best one available for that problem?
2. If no literature exists regarding the treatment method consistent with generally accepted practice?
3. Has the client been told of alternative procedures that might be preferred by the client on the basis of significant differences in disconfort, treatment time, cost, or degree of demonstrated effectiveness?
4. If a treatment procedure is publicly, legally, or professionally controversial, has formal professional consultation been obtained, has the

reaction of the affected segment of the public been adequately considered, and have the alternative treatment methods been more closely reexamined and reconsidered?

C. Is the client's participation voluntary?

  1. Have possible sources of coercion on the client's participation been considered?
  2. If treatment is legally mandated, has the available range of treatments and therapists been offered?
  3. Can the client withdraw from treatment with a penalty or financial loss that exceeds actual clinical costs?

D. When another person or an agency is empowered to arrange for therapy, have the interests of the subordinated client been sufficiently considered?

  1. Has the suborinated client been informed of the treatment objectives and participated in the choice of treatment procedures?
  2. Where the subordinated client's competence to decide is limited, have the client as well as the guardian participated in the treatment discussions to the extent that the client's abilities permit?
  3. If the interests of the subordinated person and the superordinate persons or agency conflict, have attempts been made to reduce the conflict by dealing with both interests?

Table 9.2 (Cont.)

E. Has the adequacy of treatment been evaluated?

  1. Have quantitative measures of the problem and its progress been obtained?
  2. Have the measures of the problem and its progress been made available to the client during the treatment?

F. Has the confidentialty of the treatment relationship been protected?

  1. Has the client been told who has access to the records?
  2. Are records available only to authorized persons?

G. Does the therapist refer the clients to other therapists when necessary?

  1. If treatment is unsuccessful, is the client referred to other therapists?
  2. Has the client been told that if dissatisfied with the treatment, referral will be made?

H. Is the therapist qualified to provide treatment?

  1. Has the therapist had training or experience in treating problems like the client's?
  2. If deficits exist in the therapist's qualifications, has the client been informed?
  3. If the therapist is not adequately qualified, is the client referred to other therapists, or has supervision by a qualified therapist been provided? Is the client informed of the supervisory relation?

Table 9.2 (Cont.)

4. If the treatment is administered by mediators, have the mediators been adequately supervised by a qualified therapist?

Source: Association for Advancement of Behavior Therapy. Ethical issues for human services. *Behavior Therapy*, 1977, *8*, v-vi.

light of bias in treatment of personality and behavior disturbance (Reynolds, 1981). In this regard the guidelines could have direct relevance in the assessment treatment process by providing conceptual guidelines.

The American Psychological Association,[16] American Educational Research Association,[17] and the National Council on Measurement in Education

The APA has been actively involved in providing standards for psyschologists in academic and applied settings. An early effort to address issues relating to assessment of minority children occurred within the Society for the Study of Social Issues (SSSI), Division 9 of the APA. The SSSI published a monograph in which testing of minority groups was discussed within the context of selection, use, interpretation, and sensitivity to whether or not tests differentiate reliably, validity, and are adequately interpreted with minority groups children (Deutsch, Fishman, Kogan, North, & Whiteman, 1964).

As noted above, another document prepared at the request of the APA's Board of Scientific Affairs, and entitled "Educational Uses of Tests with Disadvantaged Students" (Cleary et al., 1975) addressed several issues in testing practices: (1) it presented a review of definitions of abilities with special reference to general intelligence, (2) it summarized some common classes of test misuse and misinterpretation, (3) it reviewed the various kinds of statistical information needed to use a test effectively, and (4) discussed existing alternatives to ability tests and reviewed new types of tests and new information needed to make more effective evaluations of students in schools.

The APA also developed various standards which have a direct bearing on the assessment of individuals. For example, the *Ethical Standards of Psychologists* (APA, 1972) and *Standards for Educational and Psychological Tests* (APA, 1974) both contain guidelines on how tests are to be used and developed. The *Standards* were first developed in 1954 (at which time they were called Technical Recommendations for Psychological Tests and Diagnostic Techniques) and were endorsed by both the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME). Subsequently, the three organizations cooperated in the development of the 1966 Standards for Educational and Psychological Tests and Manuals, followed by the 1974 *Standards*. These standards are presently being revised again.

More recently, APA has endorsed several revisions in the various documents relating to training and practice that have a direct bearing on testing/assessment practices. These documents include the *Accreditation handbook* (APA, 1980), *Standards for Providers of Psychological Services* (APA, 1981b), and the *Ethical Principles of Psycholgists* (APA, 1981a).

The *Accreditation handbook* emphasizes two issues

(Pryzwansky, 1982). First, the accreditation procedures and various criteria are designed to govern accreditation of doctoral level professional psychology programs , as well as predoctoral internship programs. Second, the APA accreditation promotes high quality training along a variety of criteria including (1) institutional settings, (2) cultural and individual differences, (3) training models and curricula, (4) faculty, (5) facilities, and (6) practicum and internship settings. These criteria are important within the context of training practitioners in scientific findings in assessment and treatment and are explicit with regard to the proper training of professional psychologists to be sensitive to cultural differences.

The Standards for Providers of Psychological Services (1977) provides a uniform set of standards for psychological practice. It specifies the minimally acceptable level of quality assurance and performance for providers of psychological services. The Standards (1977) are organized around four sections that relate to a general category of service delivery.

The Standards (1977) take precedence over the Specialty Guidelines (1981) which relate to practice in each of the four specialties of professional psychology (i.e., clinical, counseling, industrial organizational, and school). Each of these Specialty Guidelines is written specifically for practice in the specialty area although there is much overlap on some (e.g., clinical and school). For example, in the

Specialty Guidelines for the Delivery of Services by School Psychologists, school psychololgical services include:

> Psychological and psychoeducational evaluation and assessment of the school functioning of children and young persons. Procedures include screening, psychological and educational tests (particularly individual psychological tests of intellectual functioning, cognitive development, affective behavior, and neuropsychological evaluations, with explicit regard for the context and setting in which the professional judgments based on assessment, diagnosis, and evaluation will be used (p. 672).

Ethical principles of psychologists (1981) provides 10 ethical principles in the areas of responsibility, competence, moral and legal standards, public statements, confidentialty, welfare of the consumer, professional relationships, assessment techniques, research with human participants, and care and use of animals.

While the Ethical Principles of Psychologists (1981) contains material relating to the psychologists' general practice, the Standards for Educational and Psychological Tests expands on these by providing more detailed and specific guidelines for test developers and users. These guidelines apply to any assessment procedure, device, or aid--i.e., to any systematic basis for drawing inferences about people (p. 2). Although these standards do not specifically deal with the concept of test bias, it can be assumed that adherence to them will reduce bias in the assessment process. However, a footnote in the Standards indicates a formal position against any testing moratorium.

Miscellaneous Professional Associations

A number of professional groups have developed some ethical guidelines for practice. It is likely that members of some of these groups would have contact with students in some sort of formal or informal assessment role. For example, the American Psychiatric Association (APA) is one of the oldest professional societies and each member is bound by the ethical code of the medical profession as defined by the Principles of Medical Ethics of the American Medical Association.

Another group that has developed a code of ethics is the National Association of Social Workers (NASW). This group, established in 1955, has over 80,000 members, many of which work in schools or with school-age children. The NASW code contains a preamble and six major sections. These sections address standards of personal and professional conduct and responsibilities to clients, colleagues, employees, and society at large.

Both the APA and NASW codes are discussed in more detail in The AAAS Professional Ethics Project (Chalk, et al., 1980).

Summary and Conclusions

In this chapter we have provided an overview of professional groups and organizations that have developed some statement or code relevant for assessment or treatment practices. It is clear that the various professional groups differ widely on positions regarding testing/asssesment practices as well as the guidelines developed therefrom. In some cases, statements of policy from one group (e.g., APA) have been directly criticized by another (ABP). Positions on both sides of the coin are often not based on empirical data. Jensen (1980) has even identified an "anti-test syndrome" with several features:

1. Most critics of tests are indiscriminate in their criticisms.
2. To most test critics there is a mystique about the word intelligence and a humanistic conviction that the most important human attributes cannot be measured or dealt with quantitatively or even understood in any scientifically meaningful sense.
3. Critics give no empirical basis for their criticisms of tests, test items, or the uses of tests.
4. Critics fail to suggest alternatives to tests-or ways of improving mental measurement-or to come to grips with the problems of educational and personnel selection or the diagnosis of problems in school learning.

5. Critics hardly ever mention the nonverbal and nonscholastic types of mental tests. They inculcate the notion that all intelligence tests simply tap word knowledge, bookish information, and use of "good English."
6. Finally, criticisms are imbued with a sense of outrage at purported social injustices either caused or reinforced by tests.

Whether or not professional organizations will be able to adopt an empirical perspective on tests in the future remains to be determined.

A final issue concerns the relationship between various professional organizations and the Supreme Court. Although many professional organizations have been active to influence assessment practices, large and far-reaching change in psychological and educational assessment practices with minority children in educational settings did not occur until impetus was provided by legislative and judicial sectors (e.g., PL 94-142). On the other hand, some authors have noted that the Court appears to be moving away from reliance upon and deference to federal agency guidelines and toward reliance upon professional standards (e.g., Standards for Educational and Psychological Tests, 1974). Learner (1978) raised this issue at a time when courts were getting involved in deciding on cases relevant to minority group assessment. Yet, as noted in Chapter 8 , Judge Grady considered expert testimony, but decided, on an individual basis, whether items in the WISC, WPPSI, WISC-R, and Stanford Binet were culturally biased against black children. Thus, at this time it is not at all clear whether or not courts in general or the Supreme Court will rely more on professional associations in rendering decisions on testing/assessment issues.

Chapter 10

Current Status

Psychological and educational assessment practices continue to increase in popularity. In public schools alone, it is estimated that about one quarter of a million standardized tests are administered annually, many for special education decision-making (Ysseldyke & Algozzine, 1982). In making decisions about special education classification and placement, test data are often collected without clear purpose and in ways not intended by their developers (Ysseldyke, Algozzine & Thurlow, 1980). This problem becomes more acute when concerns for bias in the process are addressed. The literature on bias has burgeoned in the last decade with divergent research efforts examining various aspects of the problem. While some definite progress in our understanding of bias in psychological and educational assessment reported in some areas, progress in other areas has been slow. As a consequence, there is much confusion in practice regarding methods to reduce or eliminate bias.

It was the purpose of this project to review the various aspects of the problem of bias in assessment and report our findings within a conceptual framework that provides organization to the literature. In this last chapter we examine the various parts as they contribute to the whole and make recommendations for continued research. In addition, we report on the implications of what we now know about bias for special education decision-making and discuss guidelines that can be employed given our present understandings.

Definition of bias

As a result of our review, it is apparent that a consensus definition of bias needs to be adopted. This definition should be broad enough to encompass all legitimate perspectives yet restricted in the sense that it makes no premature assumptions that can otherwise be empirically investigated. In addition, the definition should allow for the determination of bias on empirical grounds. Given these parameters, our proposed definition of bias excludes the notion that there should be an a priori assumption that mean scores, or distributions of scores should be similar across groups. This assumption denies the possibility that differences across groups are real differences. A denial of this possibility may result in time and money being invested in the development of tests and procedures that have less utility than those currently employed.

The requirement that the definition allows for the determination of bias on empirical grounds excludes those considerations that require value judgments on the part of the decision-maker(s). The "social good" or "social evil" that results from the process, while essential to consider, are presently conceived as issues of fairness and not bias. Such a distinction allows developers of tests and assessment strategies to employ common criteria to examine bias. It is the responsibility of those who employ tests to determine fairness of the strategies in the situation they intend to use them. It is the responsibility of test developers to make the distinction between bias and fairness clear, and offer guidelines that can be used by test consumers in evaluating whether or not the instrument is being employed in a manner that is fair from the consumers point of view.

This lack of a clear distinction in the past has resulted in no one willing to accept responsibility for questionable practices. Test developers have argued that their responsibility is to develop valid and reliable tests while test consumers have argued that they are only employing tests as described by test developers. This situation can be avoided by clearly articulating the distinction between the concepts of bias and fairness and identifying who is responsibile for each. National Associations can help by developing guidelines for its members in proper test development and use. Such guidelines would also prove valuable to the courts in examining issues of bias and fairness.

In keeping with the majority opinion as we believe it to be, bias is presently defined through the use of the concept of validity. It maintains the notions of bias expressed in the technical test bias and situational bias literatures and expands them to include our conceptualization of outcome bias. From this perspective, bias is present if: 1) there are differences across groups as commonly studied within the context of content, construct, and predictive validity, 2) situations and circumstances in which the assessment strategy is employed results in differences in the maximal performance across groups, and 3) their use results in difference across groups in the effectiveness of outcomes predicted from their use. Note that this latter statement expands the study of bias to all data employed in decision-making and not just test data.

Jensen (1980) reports that bias "refers to systematic errors in the *predictive validity* or the *construct validity* of test scores of individuals that are associated with the individual's group membership" (p. 375). To this definition we add the notion of outcome bias, expand it to include all forms of assessment data, rename the traditional concepts of predictive and construct validity, external and internal construct bias, respectively, and incl in our understanding of internal construct bias the notion of situational bias. The resulting definitions reads: *Bias refers to systematic errors in the internal and external construct validity or the outcome validity of tests and assessment strategies that are associated with the individual's group membership*.

Research Needs

As reported in Chapters 4, there has been substantial progress made in our study of bias from a traditional psychometric perspective. This literature suggests that little to no evidence of bias is found in commonly employed measures cognitive functioning. Research in the area of internal construct bias has shown that these tests appear to be measuring the same construct across groups with a high degree of accuracy. In addition, there does not appear to be any particular types of item that results in systematic error across groups although more research in this area utilizing latent trait statistics is needed (Cole, 1981). Yet, even if biasing items are found, their elimination is not likely to significantly contribute to a decrease in the differences now found in performance across groups.

Research investigating group differences in performance on items or clusters of items provides information on whether or not the same constructs are being measured for all. It does not tell us to what degree the construct is being measured across groups. In Chapter 5, we examined some of the situational factors that have been investigated to determine if differences across groups are related to differences in the degree to which the construct is measured. Research in this area has demonstrated that test scores can be manipulated by varying situational factors. Consequently, it does not appear that tests of cognitive functioning are measuring maximal performance. The issue of whether or not there are differences in the degree to which these situational factors can influence performance across groups is not yet clear. More research is needed to determine the differential effect of situational factors across groups.

In the study of external construct bias, psychological tests commonly employed in decision-making have been shown not to differ in their prediction of external criteria or overpredict the performance of minority group members when a nonminority or common regression line is employed. While these findings have been replicated across a variety of tests the external criteria that has been employed in externally validating intelligence tests has been criticized. These studies have commonly employed standardized measures of achievement, measures that are considered by some to be measuring the same thing as intelligence tests. Since learning is the major criteria to which intelligence tests are suppose to predict, research is needed that employs various criteria of learning to externally validate these tests and study bias.

As reported in Chapter 6, the area in most need of research is outcome bias. Much of the research efforts to date have focused on bias in measuring constructs, not in the use of tests in predicting outcomes. This appears to be a function of the way in which validity has been defined. Indeed, not only is there little research with respect to outcome bias, the research evidence regarding outcome validity, especially as it relates to intervention planning, has been limited. Research efforts specific to selection decisions have, for the most part, been limited to selection in employment. The few research efforts that have studied validity in selecting children in need of special help in schools have been hampered by methodological problems. While it appears that the predictive utility of IQ tests is reduced when measures of school achievement as opposed to academic achievement are employed, the research is equivocal with respect to bias in predicting school achievement. This would be a fruitful area to research.

With respect to the validity of tests for intervention planning and consequent bias as a result of this planning, much research is still needed. As described in Chapter 6, those assessment strategies that directly measure the behaviors of concern and are employed continuously throughout the intervention have been the only type to offer empirical evidence of outcome validity. No research reporting on intervention bias was found.

Certainly, the notion of outcome validity is the most controversial aspect of our definition of bias. There are those who maintain that the purpose of tests are to measure constructs and the validity and consequent bias of tests needs to be studied separate from the specific use of the test. The argument continues that if tests are employed inappropriately and, for example, children are selected inappropriately or interventions planned that are ineffective or biased, then the fault is not with the test. Its job is to measure the construct and should be judged on its ability to do so. While this argument has its appeal, we see certain problems with it. Basically, we question the employment of constructs unless there is evidence that their use will help in decision-making. Consequently, we have argued that evidence for their use should be part of the concepts of validity and bias.

Broadened conceptions of validity and bias connote that the measurement of the construct is not an end in itself, but only a means to an end. Indeed, the sole reason for inventing constructs is to serve some purpose. A broadened notion of validity would require that its purpose be empirically established. Consumers now turn to psychological measures, evaluate whether or not they measure the construct of interest, and then infer that its use is of value to them. The knowledgable consumer appreciates the inferences that are made and uses caution. We would prefer that they use empirical evidence instead. Such evidence, we believe, can best be generated if we change our understanding of validity to incorporate outcome valdity.

As discussed in the beginning of this chapter, a differentiation needs to be made between bias and fairness. This is by no means an attempt to downplay the importance of fairness. Ultimately, all decisions regarding the employment of psychological and educational assessment strategies rests on the decision-maker's beliefs of whether or not the whole process is consistent with their values and/or the values of those who employ them. At present, the judgment regarding whether a decision-making process is fair or not is an unsystematic exercise, if an exercise at all. Guidelines are needed that can make this exercise systematic by having decision-makers think through whether or not the product of their efforts is ultimately fair to all.

The models of selection reviewed in Chapter 6 were designed to reflect the various philosophies of fairness. What is needed next is a clear exposition of the models so that they can be understood and employed by consumers. We recommend that the Expected Utilities Model be engineered for consumer use since it can be employed regardless of what fairness philosophy is adopted by decision makers. It would provide a vehicle for making whatever adjustments to decisions are necessary.

With regard to the various proposed alternatives to present practice, several seem promising for further investigation and implementation. Behavioral Assessment and Criterion-Referenced Testing appear to be the most highly developed procedures that can have an immediate impact on planning interventions to help children develop skills in which they are deficient. Strategies such as Diagnostic/Clinical Teaching and Child Development Observation still need to be researched to validate their usefulness.

For diagnostic purposes, Learning Potential Assessment may hold promise in its ability to measure learning as it occurs in assessment rather than as a static, after the fact occurance. While this may reduce the potential impact of biasing historical factors, this is an empirical question that has yet to be addressed. However, the most active proponents of this method, Fuerstein and his colleagues, have moved into using learning potential assessment as a method of identifying skills for intervention. Much research in their procedures, identified as Instrumental Enrichment, needs to be completed before the validity of its use for intervention planning is established.

The value of these strategies that are purported to improve upon the diagnostic and predictive capability of IQ tests show little evidence to support their continued development at this time. Renorming to correct for empirically established bias is a legitimate activity but to renorm a test on the a priori assumption that mean scores across groups should be similar appears questionable from our perspective. Developing culture-reduced tests also appears unwarranted. Culturally loaded items, if any are found to bias responses across groups, can be eliminated from present tests. Developing new tests with all items culture-reduced does not appear to be worth the time or expense. In short, developing new instruments for diagnosing and classificating appears to be misdirecting our efforts by overemphasizing its importance in helping children.

Guidelines for Reducing Bias in Special Education Decision-Making

Our understanding of bias and what to do about it continues to develop as we investigate it. What is it that can be done at present to help implement what we know and what cautions can be exercised in areas that are still under investigation? Tucker (1980) recommends a series of steps that can be taken with regard to special education decision making that address this question. "The list specifies nineteen points in the appraisal process at which assessment data is (or should be) collected and used in evaluating a student's program from a non-biased perspective" (Tucker, 1980, p.3). The steps are based on a series of questions that decision-makers ask themselves throughout the process that leads up to the classification and placement of children in special education classes. The nineteen questions are listed below:

(1) Is there a significant problem involving this student?

(2) Is the problem worth taking time to pursue?

(3) Does the initial observational data collected on a day to day basis suggest that a significant problem exists?

(4) Does the information gained from the parent or guardian suggest the need for alternative classroom intervention?

(5) Do the observational data from Step 4 show that the problem behavior persists even when alternative classroom strategies are implemented?

(6) Does the screening data suggest the need for other alternative educational services?

(7) Does the problem persist even when alternative regular education alternatives (sic) are provided?

(8) Have all steps 1 through 7 been taken and is all of the resulting data on hand?

(9) Have all the necessary questions been generated to provide an adequate biasis for planning the student's educational program?

(10) After the assessment performed in Step 9, is there sufficient evidence that the student is handicapped?

(11) Does the assessment data obtained in Step 10 supply sufficient evidence that the student's problem is educationally related to and supported by a handicapping condition?

(12) Have all the assessment questions been answered to the satisfaction of the Multidisciplinary Team?

(13) Is the Assessment Report jargon free and understandable in that it communicates in simple, straigtforward terms to all who will be present at the I.E.P. meeting?

(14) Does the student appear to need special education?

(15) Is the student a member of a minority group or other unique population?

(16) Are eligibility decisions free of cultural bias?

(17) Have all the necessary precautions been taken to insure that the student's educational needs can best be met by the provision of special education services?

(18) Have the parents approved the student's placement and the program as specified in the I.E.P.?

(19) Can we tell if the student's progress is satisfactory?

After each of the questions, Tucker provides discussion concerning how to proceed depending on whether the answer is "yes" or "no".

As can be seen from the questions, Tucker has basically provided guidelines for conducting a thorough assessment consistent with the requirements embodied in P.L. 94-142. Though issues specifically related to minority group assessment are not addressed in every step, it is implied that if decision-makers are thorough in their method, potential bias can be reduced. For example, question 3 may reduce biased impressions by requiring that a problem be documented by observational data. Likewise, question 5 may reduce bias in labeling and placement by employing alternative classroom strategies in an attempt to remedy the problem in a less intrusive way.

An interesting recommendation of Tucker is that the multidisciplinary team include a member that is sensitive to the student's racial or cultural group. Such a step may help in identifying potential biases in the team that occur out of ignorance of what is "expected" of a child from a certain racial or cultural group. Ultimately, the validity of all such impressions need to be empirically documented and the potential bias examined.

In preparing for a formal evaluation of minority children, Tucker recommends that teams have information on language proficiency and differences between expressive and receptive language. These recommendations are consistent with P.L. 94-142 guidelines and are justifiable based on the empirical literature. The languages of the child and the examiner do make a difference in the assessment of bilingual children. While the short-range predictive validity of IQ tests is as good for bilingual students as it is for English speaking students, this finding appears to result only because the criterion-measure also requires English proficiency. Evidence that there is a significant difference between verbal and performance IQ measures within groups as a function of language proficiency strongly suggests that the use of tests that emphasize verbal abilities are biased against bilingual students. Consequent recommendations that performance measures be used to assess intellectual functioning of bilingual students are sound.

Consistent with the law, Tucker (1980) also recommends that substandard performance on a measure of adaptive behavior be a prerequisite to further evaluation for mental retardation. This is also in keeping with current definitions of mental retardation. If the identified purpose for administering an IQ test is to diagnose retardation and the diagnosis can only be accompanied by significantly substandard performance in adaptive behavior, then IQ testing only needs to be performed with those children who remain eligible after the administration of an adaptive behavior measure. Of course, the converse of this statement is also true. That is, if the child's performance on an IQ test does not qualify him/her for classification then there is no need to administer an adaptive behavior measure. It would seem less intrusive to the child and more cost efficient if data on adaptive behavior were collected first rather than IQ data.

As reviewed in Chapter 7, certain adaptive behavior measures such as the ABIC and ABS-SE, have a degree of content and construct validity. The predictive and outcome validities of these instruments, however, have yet to be established.

The importance of establishing the validity of the use of IQ tests, adaptive behavior measures, or both in the diagnosis and placement of mentally handicapped children becomes highlighted when decisions concerning proportional representation need to be made. If proportional representation is determined to be fair by the decision makers, how does one best choose the proportions of each group to be classified/placed? As discussed previously, there are many models from which to choose in making this decision. However, in this case, all models would require that we know which data would provide the best prediction. If, for example, only 20% of an EMR population is allowed to be black and there are more than that percentage identified and placed through the use of IQ tests alone, would it be a more valid strategy to rank the children to be chosen on IQ alone or rank and choose the number by adopting a measure of adaptive behavior to combine with IQ testing. Indeed, as Fisher (1978) reports, the use of the ABIC in combination with IQ tests would reduce by 60, 70, and 85% the number of Anglo, Mexican-American, and black children, respectively, that are classified EMR by IQ tests alone. Such a strategy may in and of itself solve a district's problems. The answer, of course, has to be based on what you are predicting and the validities of each procedure in making these predictions. With respect to EMR diagnosis, the answer has already been provided. By definition, EMR diagnosis now requires both and the courts concur. Reschly (1982) points to one source of validity for using both IQ and adaptive behavior to predict classification. If both are used, the prevalence of mental retardation in the schools will drop to approximately 1 to 1.5%. This figure closely approximates the percentage of adults identified mentally retarded through community surveys (Tarjan, 1970, cited in Reschly, 1982). Whether or not the placement that always follows such diangosis has any outcome validity is not known.

The implementation of Tucker's (1980) nineteen steps are designed to address issues related to potential bias in the process. He provides no guidelines regarding fairness in selecting and placing students.

Sattler (1982) reports on several recommendations that have been made for assessing ethnic minority children. Those that are consistent with our understanding of the literature are reported below.

(1) Assessment should focus on discovering ways to help children and not on ways to better classify/place. The increased use of behavioral and criterion-referenced measures designed to intervene on specific skills deficits should be the focus of the assessment.

(2) Examiners should take the time to motivate children to perform on tests. Seeking the cooperation of children would help reduce problems reported in situation bias.

(3) A wide range of mental tests should be used when assessing minority children. Given the complex nature of intellectual functioning, the more adequate the sample of skills assessed, the more likely one is to gain a more valid and reliable measure for any one child.

(4) Procedures for "testing the limits" (Sattler, 1974, 1982) may provide a better picture of the maximal performance capabilities of a child than standardized procedures alone.

(5) Emphasize or use exlusively nonlanguage performance measures with bilingual children. Using bilingual examiners who can allow the child the opportunity to respond in the language they prefer is desirable.

(6) Clinicians should become knowledgable of the cultural and racial differences among children in the community they work.

(7) Teachers should become sensitive to the educational difficulties displayed by various minority group children. Behaviors that connote one type of problem for some groups may reflect a different problem for others.

Concluding Remarks

One of the most fundamental needs in the area of nonbiased assessment that has arisen from our review is the need for test consumers to carefully examine the purposes for which they assess and test developers to broaden their concept of validity and consequent study of bias to provide consumers with the research information they need to select the best strategies to fulfill these purposes. With respect to test consumption, when one asks the basic questions regarding the purposes of special education decision-making, one can't help but wonder why there is the need to diagnose at all. If the purpose of special education is to provide special help to children with learning problems, then we can't help but ask the same question that has been reiterated over the last few decades, why diagnose? With evidence continuing to mount regarding the superfluous nature of the activity, time for change is imminent. Such change is now being witnessed in several states and it is not radical to predict that educational classfication as it is now conceived will eventually disappear. Noncategorical special education placement based on a child's educational needs rather than his classification will hopefully prove to be the next major change in providing help to children. Such a system would focus attention on collecting data to help children rather than diagnosing them.

Once decision-makers decide on the purpose of their assessment activities they will need better data on the validity and unbiased nature of that data to make the decision they deem important. Such information can best be accomplished by expanding our understanding of the concept of validity to include outcome validity. This would then allow for an examination of selection and intervention bias.

As test consumers evaluate the purposes of their activities, an examination of their own values that ultimately impact on the decisions they make need to be determined. Such an examination of the fairness of their activities would openly address issues that are often only casually addressed or ignored.

References

Accredition Handbook. Washington, D.C.: Author, 1980.

Achenbach, T.M. Developmental psycholpathology. New York: Ronald Press, 1974.

Agras, W.S. Toward the certification of behavior therapists? Journal of Applied Behavior Analysis, 1973, 6, 167-173.

Aleviozos, P., DeRisi, W., Liberman, R., Eckman, T., & Callahan, E. The behavior observation instrument: A method of direct observation for program evaluation, Journal of Applied Behavior Analysis, 1978, 11, 243-257.

Alexander, F.M. & Selesnick, S.T. The history of psychiatry: An evaluation of psychiatric thought andpractice from prehistoric times to the present. New York: New American Library, 1968.

Algozzine, B., Mercer, D.C., & Countermine, T. The effects of labels a.d behavior on teacher expectations, Exceptional Children, 1977, 44, 131-132.

Ali, F., & Costello, J. Modification of the Peabody Picture Vocabulary Test. Developmental Psychology, 1971, 5, 86-91.

Alley, G., & Foster C. Nondiscriminatory testing of minority and e8ceptional children, Focus on Exceptional Children, 1978, 9, 1-14.

American Psychological Association. Ethical standards of psychologists. Washington, D.C.: Author, 1972.

American Psychological Association. Ethical principles in the conduct or research with human participants. Washington, D.C.: Author, 1973.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. Standards for educational and psychological tests. Washington, D.C.: American Psycholcgical Association, 1974.

American Psychological Association. Ethical principles of psychologists, American Psychologists, 1981, 36, 633-638. (a)

American Psychological Association. Specialty guidelines for the delivery of services, American Psychologist, 1981, 36, 639-681. (b)

Anastasi, A. Diverse effects of training on tests of academic intelligence. In B.G. Green (Ed.) Issues in testing: Coaching, disclosure, and ethnic bias. San Francisco: Jossey-Bass, 1981.

Anastasi, A. Psychological Testing (4th Ed.). New York: Macmillan, 1976.

Anastasi, A. Psychological tests: Uses and abuses, Teachers College Record, 1961, 62, 38-393.

Anastasi, A. Differential Psychology (3rd ed.). New York: Macmillan, 1958.

Anastasi, A., & Cordova, F.A. Some effects of bilingualism upon the intelligence test performance of Puerto Rican children in New York, Journal of Educational Psychology, 1953, 44, 1-19.

Anderson, T., Cancelli, A.A., & Kratochwill, T.R. Self-reported assessment practices of school psychologists, *Journal of School Psychology*, in press.

Angoff, W.H., & Ford, S.G. Item-race interaction on a test of scholastic aptitude, *Journal of Educational Measurement*, 1973, *10*, 95-106.

Arbitman-Smith, R. & Haywood, G.C. Cognitive education for learning-disabled adolescents, *Journal of Abnormal Child Psychology*, 1980, *8*, 51-64.

Arena, J.J. *Teaching through sensory-motor experiences*. San Rafael, Ca.: Academic Therapy Publications, 1969.

Arieti, S. Psychiatric controversy: Man's ethical dilemma, *American Journal of Psychiatry*, 1974, *132*. 763-764.

Arvey, R.D. Some comments on culture fair tests, *Personnel Psychology*, 1972, *25*, 433-448.

Association for Measurement and Evaluation in Guidance, American Personnel and Guidance Assocciation & National Council for Measurement in Education. The responsible use of tests: A position paper of AMEG, APGA, and NCME. *Measurement and Evaluation in Guidance*, 1972, *5*, 385-388.

Ayllon, T., & Kelly, K. Effects of reinforcement on standardize$ test performance, *Journal of Applied Behavior Analysis*. 1972, *5*, 477-484.

Ayres, A.J. Sensorimotor foundations of academic ability. In W.M. Cruickshank & D.P. Hallahan (Eds.), Perception and learning disabilities in children. Vol. 2. Syracuse, New York: Syracuse University Press, 1975.

Baer, D.M. Perhaps it would be better not to know everything, Journal of Applied Behavior Analysis, 1977, 10, 167-172.

Baer, D.M., Wolf, M.M., & Risley, T.R. Some current dimensions of applied behavior analysis, Journal of Applied Behavior Analysis, 1968, 1, 91-97.

Bandura, A. Social learning theory. Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change, Psychological Review, 1977, 84, 191-215. (b)

Bandura, A. Psychotherapy based upon modeling principles. In A.E. Bergin and S.L. Garfield (Eds.), Handbook of psychotherapy and behavior change. New York: John Wiley, 1971.

Bandura, A. Principles of behavior modification. New York: Holt, Rinehart, & Winston, 1969.

Bandura, A. & Walters, R.H. Social learning and personality development. New York: Holt, Rinehart & Winston, 1963.

Barker, G.R., Dembo, T., & Lewin, K. Frustration and aggression. In R.G. Barker, J.S. Kounin, & H.T. Wright (Eds.), ChilS Behavior and Development. Stanford, California: Stanford University Press, 1968.

Barker, R.G., & Wright, H.F. Psychological ecology and the problem of psychosocial development, *Child Development*, 1949, *20*, 131-143.

Barker, R.G., & Wright, H.F. *Midwest and its children*. Evanston, Illinois: Row, Peterson & Co., 1954.

Barlow, D.H. (Ed.) *Behavioral assessment of adult disorders*. New York: Grulford Press, 1981. (a)

Barlow, D.H. On the relation of clinical research to clinical practice: Current issues, new directions, *Journal of Connsulting and Clinical Psychology*, 1981, *49*. 147-155. (b)

Barlow, D.H. Behavior therapy: The next decade, *Behavior Therapy*, 1980, *11*, 315-328.

Barlow, D.H., & Mavissakalian, M. Directions in the assessment and treatment of phobia: The next decade. In M. Mavissakalian & D.H. Barlow (Eds.), *Phobia: Psychological and pharmacological treatment*. New York: Guildford Press, 1981.

Barlow, D.H. & Wolfe, B.E. Behavioral approaches to anxiety disorders: A report on the NIMH-SUNY, Albany Research Conference, *Journal of Consulting and Clinical Psychology*, 1981, *49*, 448-454.

Barsch, R.H. *Enriching perception and cognition*. Seattle, Wash.: Special Child Publications, 1968.

Barsch, R.H. *Achieving perceptual-motor efficiency*. Seattle, Wash.: Special Child Publications, 1967.

Barsch, R.H. *A movigenic curriculum*. Madison, Wis.: Bureau for Handicapped Children, 1965.

Bartlett, C.H., Bobko, P., Mosier, S.B., & Hannan, R. Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis, *Personnel Psychology*, 1978, *31*, 233-241.

Bartell, N., Grill, J., & Bryen D. Language characterisitcs of black children: Implications for assessment, *Journal of School Psychology*, 1973, *11*, 351-364.

Bateman, B. An educator's view of a diagnostic approach to learning disorders. In J. Hellmuth (Ed.), *Learning Disorders*. Seattle: Special Child Publications, 1965, *1*, 219-239.

Bateman, B.D., & Schiefelbusch, R.L. Educational identification, assessment, and evaluation procedures. In *Minimal brain dysfunction in children* (Phase II). N & SDCP Monograph, U.S. Department of Health, Education, and Welfare, 1969.

Baughman, E.E., & Dahlstrom, W.G. *Negro and white children: A psychological study in the rural south*. New York: Academic Press, 1968.

Beberfall, L. Some linguistic problems of the Spanish-speaking people of Texas, *Modern Language Journal*, 158, *42*, 87-90.

Becker, H. *Outsiders: Studies in the sociology of deviance*. New York: Free Press, 1963.

Bee, H.L., Streissguth, A.P., Van Egevan, L.F., Leckie, M.S., & Nyman, B.A. Deficits and value judgments: A comment of Sroufe's critique. *Developmental Psychology*, 1970, *2*, 126-149.

Behrmann, P. Activities for developing visual perception. San Rafael, Ca.: Academic Therapy Publications, 1970.

Bellack, A.S., & Hersen, M. Assessment and single-case research. In M. Hersen & A.S. Bellack (Eds.) Behavior therapy in the psychiatric setting. Baltimore: The Williams & Wilkins Co., 1978.

Bellack, A.S., & Hersen, M. Behavior modification: An introductory textbook. Baltimore: Williams & Wilkins, 1977. (a)

Bellack, A.S., & Hersen, M. Self-report inventories in behavioral assessment. In J.D. Cone and R.P. Hawkins (Eds.) Behavioral assessment: New directions in clinical psychology. New York: Brunner/Mazel, 1977. (b)

Benjamin, L.S. Structural analysis of social behavior. Psychological Review, 1974, 81, 392-425.

Benton, A.L. Influence of incentives upon intelligence test scores of school children, Journal of Genetic Psychology, 1936, 49. 494-497.

Bergan, A., McManis, D.L., & Melchert, P.A. Effects of social and token reinforcement on WISC block design performance, Perceptual Motor Skills, 1971, 32, 871-880.

Bergan, J.R. Behavioral consultation. Columbus, Ohio: Charles E. Merrill, 1977.

Bergan, J.R., & Tombari, M.L. Consultant skill and efficiency and the implementation and outcomes of consultation. Journal of School Psychology, 1976, 14, 3-14.

Bergan, J.R. & Tombari, M.L. The analysis of verbal interactions occurring during consultation, Journal of School Psychology, 1975, 13, 209-226.

Bernetta, M. Visual readiness and developmental visual perecption for reading. Journal of Developmental Reading, 1962, 5, 82-86.

Bersoff, D.N. The legal regulation of school spychology. In C.R. Reynolds & T.B. Gutkin (Eds.), Handbook of School Psychology. New York: Wiley, 1981.

Bersoff, D.N. Testing and the law. American Psychologist, 1981, 36, 1047-1056.

Bersoff, D.N. Legal and ethical concern in research. In L. Goldman (Ed.) Research methods for counselors. New York: Wiley, 1978.

Bersoff, D.N. Silk purses into sows' ears: The decline of psychological testing and a suggestion for its redemption. American Psychologist, 1973, 28, 892-899. (a)

Bersoff, D.N., Larry P. and PASE: Judicial report cards on the validity of individual intelligence tests. In T.R. Kratochwill (Ed.) Advances in school psychology (Vol. 2). Hillsdale, N.J.: Lawrence Erlbaum, 1982.

Bijou, S.W. Child development: The basic stage of early childhood. Englewood Cliffs, New Jersey: Prentice-Hall, 1976.

Bijou, S.W. The critical need for methodological consistency in field and laboratory studies. Paper given at the First Symposium of the International Society for the Study of Behavior Development. Nijmegen, July, 1971.

Bijou, S.W. What psychology has to offer education--now. Journal of Applied Behavior Analysis, 1970, 3, 65-71.

Bijou, S.W., & Baer, D.M. Child development: Universal stage of infancy (Vol. 2). Englewood Cliffs, N.J.: Prentice-Hall, 1965.

Bijou, S.W., & Grimm, J.A. Behavioral diagnosis and assessment in teaching young handicapped children. In T. Thompson & W.S. Dockens III (Eds.), Applications of behavior modification. New York: Academic Press, 1975.

Bijou, S.W., & Peterson, R.F. Psychological assessment in children: A functional analysis. In R. McReynolds (Ed.), Advances in psychological assessment (Vol.2). Palo Alto, Calif.: Science and Behavior Books, 1971.

Bijou, S.W. Peterson, R.F., & Ault, M.H. A method to integrate descriptive and experimental field studies at the level of data and empirical concepts, Journal of Applied Behavior Analysis, 1968, 1, 175-191.

Bijou, S.W., Peterson, R.F., Harris, F.R., Allen, K.E., & Johnson, M.S. Methodology for experimental studies of young children in natural settins, Psychological Record, 1969, 19, 177-210.

Binet, A., & Henri, V. La psychologie individuelle. Annee Psychologique, 1895, 2, 411-463.

Bingham, W.V., Moore, B.V., & Gustad, J. *How to interview*. (4th Ed.) New York: Harper, 1959.

Bisett, B.M., & Reiber, M. The effects of age and incentive value on discrimination learning, *Journal of Experimental Child Psychology*, 1966, *3*, 199-206.

Black, H. *They shall not pass*. New York: Morrow, 1963.

Blatt, B. The physical personality and acadmic status of children who are mentally retarded attending special classes as compared to children who arementally retarded attending regular classes. *American Journal of Mental Deficiency*. 1958, *62*, 810-818.

Block, M.J., & Dworkin, G. (Eds.) *The IQ controversy: critical readings*. New York: Pentheon, 1976.

Blum, A. Sociology of mental illness. In J.D. Douglas, (Ed.), *Deviance and respectability: The social construction of moral meanings*. New York: Basic Books, 1970, 31-60.

Boehm, S., & Weinberg, R.A. *The classroom observer: A guide for developing observation skills*. New York: Teachers College Press, 1977.

Boehm, V.R. Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. *Journal of Applied Psychology*, 1972, *56*, 33-39.

Boorstein, D. *The Americans: The democratic experience*. New York: Vintage, 1974.

Boring, E.G *A History of Experimental Psychology* (2nd ed.). New York: Appelton-Century-Crofts, 1950.

Bosma, B. The NEA testing moratorium. *Journal of Psychology*, 1973, *11*, 304-306.

Bracht, G. Experimental factors related to aptitude-treatment interactions. *Review of Educational Research*, 1970, *40*, 627-645.

Bradfield, R.H., Brown, J., Kaplan, P., Richert, E., & Stannard, R. The special child in the regular classroom. *Exceptional Children*, 1973, *39*, 384-390.

Brady, J.P. Perspective on research with human subjects, *The Behavior Therapist*, 1979, *2*, 9-13.

Brainerd, C.J. *Piaget's theo2y of intelligence*. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Braue, B.B., & Masling, J.M. Intelligence tests used with special groups of children, *Journal of Exceptional Children*, 1959, *25*, 42-45.

Braun, S.H. Ethical issues in behavior modification, *Behavior Therapy*, 1975, *6*, 51-62.

Brigham, C. *A Study of American Intelligence*. Princeton, N.J.: Princeton University Press, 1923.

Brigham, C. Intelligence tests of immigrant groups, *Psychological Review XXXVII*, 158-165, 1940.

Brigham, T.A., Graubard, P.S. & Stans, A. Analysis of the effects of sequential reinforcement contingencies on aspects of composition. *Journal of Applied Behavior Analysis*, 1972, *5*, 421-429.

Brody, G.H., & Brody, J.A. Vicarious language instruction with bilingual children through self-modeling. Contemporary Educational Psychology, 1976, 1, 138-145.

Brown, D.K., Kratochwill, T.R. & Bergan, J.R. Teaching interview skills for problem identification: An analogue study, Behavioral Assessment. 1982, 4, 63-73.

Brown, F. The SOMPA: A system of measuring potential abilities? School Psychology Digest. 1979, 8, 37-46.

Brown, L., Nictupski, J. & Hamre-Nietupski, S. Criterion of ultimate functioning. In M.A. Thomas, Hey don't forget about me! Reston, VA: Council for Exceptional Children, 1976.

Browning, R.M., & Stover, D.O. Behavior modification in child treatment: An experimental and clinical approach. Chicago: Aldine-Atherton, 1971.

Brunswick, E. Systematic and representative design of psychological experiment. Berkeley: Cal. University of California Press, 1947.

Buckley, K., & Oakland, T. Contrasting localized norms for Mexican-American children on the ABIC. Paper presented at the annual meeting of the American Psychological Association, San Francisco, August, 1977.

Bucky, S., & Banta, T. Racial factors in test performance, Developmental Psychology. 1972, 6, 7-13.

Budoff, M. Providing special education without special classes, Journal of School Psychology. 1972, 10, 199-205.

Budoff, M. Learning potential: A supplementary procedure for assessing the ability to reason. Seminars in Psychiatry, 1969, 1, 278-290.

Budoff, M. Learning potential among institutionalized young adult retardates, American Journal of Mental Deficiency. 1967, 72, 404-411.

Budoff, M., & Friedman, M. "Learning potential" as an assessment approach to the adolescent mentally retarded. Journal of Consulting Psychology, 1964, 28, 433-439.

Budoff, M., & Hutton, L. The development of a learning potential measure based on Raven's Professive Matrices, Studies in Learning Potential, 1972, 1, 18.

Budoff, M., Meskin, J. & Harrison, R. Educational test of the learning-potential hypothesis, American Journal of Mental Deficiency, 1971, 76, 159-169.

Buros, O.K. (Ed.) Mental Measurements Yearbook (7 Vols.) Highland Park, N.J.: Gryphon Press, 1938-1972.

Callender, J.C, & Osburn, H.G. Development and test of a new model of validity generalization, Journal of Applied Psychology, 1980, 65, 543-558.

Cancelli, A.A. Behavioral assessment: Structure on domains in psychoedu#ational assessment. In J.R. Bergan (Chair), Behavior approaches to psychoeducational assessment. Symposium presented at the meeting of the American Psychological Association, Toronto, Canada, August, 1978.

Cancelli, A.A. & Duley, S.M. Behavioral assessment in school psychology. In J.R. Bergan (Ed.) *Contemporary school psychology*. Columbus, Ohio: Merrill, in press.

Cancelli, A.A. & Kraotchwill, T.R. (Ed.) *Advances in School Psychology* (Vol 1) Hillsdale, N.J.: Lawrence Erlbaum, 1981.

Cardall, C., & Coffman, W.E. *A method for comparing the performance of different groups on the items of a test*.

Carmine, D., & Silbert, J. *Direct instruction reading*. Columbus, Ohio: Charles E. Merrill, 1979.

Carroll, J.B. On the theory-practice interface in the measurement of intellectual abilites. In Patrick Suppes (Ed.) *Impact of research on education: Some case studies*. Washington, D.C.: National Academy of Education, 1978.

Cartwright, C., & Cartwright, G.P. *Developing observation skills*. New York: McGraw-Hill, 1974.

Cassidy, V.M., & Stanton, J.E. *An investigation of factors involved in educational placementof mentally retarded children* (Cooperative Research Project 043). Washington, D.C.: U.S. Office of Education, 1959.

Catterall, C.D. Special education in transition: Implications for school psycholoy. *Journal of School Psychology*, 1972, *10*, 91-98.

Centra, J.A., Linn, R.L., & Parry, M.E. Academic growth in predominantly Negro and predominantly white colleges. *American Educational Research Journal*, 1970, *1*. 83-98.

Chalk, R., Rankel, M.S., & Chafer, S.B. (Eds.) *AAAS professional ethics project: Professional ethics activities in the scientific and engineering societies*. Washington, D.C.: American Association for the Advancement of Science, 1980.

Chandler, J.T., & Plakos, J. Spanish-speaking pupils classified as educable mentally retarded. *Integrated Education*, 1969, *1*, 8-33.

Chapman, P. Lewis M. Terman and the Intelligence Testing Movement, 1890-1930. Unpublished Ph.D. dissertation, Stanford University, 1979.

Chapman, M., & Hill, R.A. (Eds.) *Achievement motivation: An analysis of the literature*. Philadelphia: Research for Better Schools, Inc., 1971.

Chase, A. *The legacy of Malthus: The social costs of the new scientific racism*. New York: Alfred A. Knopf, 1977.

Chavez, S.J. Preserve their language heritage. *Childhood Education*. 1956, *33*, 165-185.

Ciminero, A.R. Behavioral assessment: An overview. In A.R. Ciminero, K.S. Calhoun and H.E. Adams (Eds.) *Handbook of behavioral assessment*. New York: Wiley-Interscience, 1977.

Ciminero, A.R., Calhoun, K.S., & Adams, H.E. (Eds.) *Handbook of Behavioral Assessment*. New York: Wiley, 1977.

Ciminero, A.R., & Draburan, R.S. Current developments in the behavioral assessment of children. In B.B. Lahey & A.E. Kazdin (Eds.), *Advances in child clinical psychology*. (Vol. 1), New York: Plenum Press, 1977.

Ciminero, A.R., Nelson, R.O., & Lipinski, D.P. Self-monitoring procedures. In A.R. Ciminero, K.S. Calhoun, & H.E. Adams (Eds.) *Handbook of Behavioral Assessment*. New York: Wiley-Interscience, 1977.

Clarizio, H.F. In defense of the IQ test. *School Psychology Digest*, 1979, *8*, 79-88. (b)

Cleary, T. A. Text bias: Predication of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 1968, *5*, 115-124.

Cleary, T.A., Humphreys, L.G., Kendrick, S.A., & Wesman, A. Educational uses of tests with disadvantaged students. *American Psychologist*, 1975, *30*, 15-41.

Clinman, J., & Fowler, R.L. The effects of primary reward on the IQ performance of grade-school children as a function of initial IQ level. *Journal of Applied Behavior Analysis*, 1976, *9*, 19-23.

Cohen, L. The effects of material and non-material reinforcement upon performance of the WISC Block Design subtest by children of different social classes: A follow-up study. *Psychology*, 1970, *4*, 41-47.

Cohen, R.A. Conceptual styles, culture conflict and nonverbal tests of intelligence. *American Anthropologist*, 1969, *71*, 828-856.

Cole, L.J. *Adaptive behavior of the educable mentally retarded child in the home and school environment*. Unpublished doctoral dissertation, University of California, Berkeley, 1976.

Cole, M., Gay, J., Glick, A. A., & Sharpe, D. W. *The cultural contest of learning and thinking*. New York: Basic Books, 1971.

Cole, N.S. Bias in testing. *American Psychologist*, 1981, *36*, 1067-1077.

Coleman, H.M. Visual perception and reading dysfunction. *Journal of Learning Disabilities*, 1968, *1*, 116-123.

Coleman, H.M., & Dawson, S.T. Educational evaluation of visual-perceptual-motor dysfunction. *Journal of Learning Disabilities*, 1969, *2*, 242-251.

Cone, J.D. Psychometric considerations. In M. Husen and H.S. Bellock (Eds.), *Behavioral Assessment: A practical handbook*. 2nd Ed. New York: Pergamon, 1981.

Cone, J.D. The behavioral assessment grid (BAG): A conceptual framework and a taxonomy. *Behavior Therapy*, 1978, *9*, 882-888.

Cone, J.D. The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, 1977, *3*, 411-426.

Cone, J.D. *Multitrait-multimethod matrices in behavioral assessment*. Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1976.

Cone, J.D. *What's relevant about reliability and validity for behavioral assessment?* Paper presented at the meeting of the American Psychological Association, Chicago, September 1975.

Cone, J.D., & Hawkins, R.P. (Eds.) *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.

Conner, J.J., & Weiss, F.L. A brief discussion of the efficacy of raising standardized tes4 scores by contingent reinforcement, *Journal of Applied Behavior Analysis*, 1974, *7*, 351-352.

Cooley, W.R., and Lohnes, P.R. *Evaluation research in education: Theory, principles and practice*. New York: Irving Publishers.

Cooper, H.M. Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, 1982, *52*, 291-302.

Costello, J. Effects of pretesting and examiner characteristics on test performance of young disadvantaged children. *Proceedings of the 78th Annual Convention of the American Psychological Association*,1970,*5*,309-310.Coulter, W.A., & Morrow, H.W. *Adaptive behavior: Concepts & measurements*. New York, Grune & Stratton, 1978.

Coulter, W.A., & Morrow, H.W. *The concept and measurement of adaptive behavior*. New York, Grune & Stratton, 1978.

Cronbach, L.J. Validity on parole: How can we go straight? In W.B. Schrader (Ed.), *New directions for testing and measurement: No. 5 Measuring achievement: Progress over a decade*. San Francisco: Jossey-Bass, 1980.

Cronbach, L.J. Five decades of public controversy over mental testing, *American Psychologist*, 1975, *30*, 1-14.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.). *Educational measurement* (2nd Ed.). Washington, D.C.: American Council on Education, 1971.

Cronbach, L.J. *Essentials of psychological testing* (2nd Ed.). New York: Harper & Row, 1960.

Cronbach, L.J., Gleser, G.C., Nada, H., & Rajaratman, N. *The dependability of behavioral measures*. New York: Wiley, 1972.

Cronbach, L.J., & Snow, R.E. *Aptitudes and instructional methods: A handbook for research in interactions*. New York: Irvington/Maiburg, 1976.

Crown, P.J. *The effects of race of examiner and standard vs. dialect administration of the Wechsler Preschool and Primary Scale of Intelligence on the performance of Negro and white children*. Doctoral dissertation, Florida State University, Ann Arbor, Mich.: University Microfilms, 1970, No. 71-18, 356.

Darlington, R.B. Another look at "culture fairness". *Journal of Educational Measurement*, 1971, *8*, 71-82.

Davis, J.A., & Kerner-Hoeg, S. *Validity of preadmissions indices for blacks and whites in six traditionally white public universities in North Carolina*. ETS Report PR-71-15. Princeton, N.J.: Educational Testing Service, 1971.

Davis, J.A., & Temp, G. Is the SAT biased against black students. *College Board Review*, Fall 1971, pp. 4-9.

Davison, L.A. Introduction to clinical neuropsychology. In Reitan, R.M., & Davison, L.A. (Eds.) *Clinical neuropsychology: Current status and applications*. Washington, D.C.: Winston, 1974.

Day, W.F. Contemporary behaviorism and the concept of intention. In M.R. Jones (Ed.) Nebraska symposium on motivation (Vol. 23). Lincoln: University of Nebraska Press, 1976.

Dayton, C.M., & Macready, G.B. A probablisitic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

DeAvila, E.A., & Havassy, B.E. Piagetian alternatives to IQ: Mexican American study. In N. Hobbs (Ed.) Issues in the classification of exceptional children. San Francisco: Jossey-Bass, 1975.

DeAvila, E.A., & Havassy, B.E. Intelligence of Mexican American children: A field study comparing neoPiagetian and traditional capacity and achievement measures. Austin, Texas: Dissemination Center for Bilingual Bicultural Education, 1974.

deHirsch, K., Jansky, J.J. & Lanford, W.S. Predicting reading failure. New York: Harper & Row, 1966.

Des Jarlais, D.C. Mental illness as social deviance. In W.C. Rhodes and M.L. Tracy (Eds.) A study of child variance Volume I: Conceptual models. Ann Arbor: The University of Michigan, 1972.

Deno, S. The Seward University Project: A cooperative effort to improve school services and university training. In E. Deno (Ed.), Intructional alternatives for exceptional children. Arlington, Va.: Council for Exceptional Children, 1973.

Deno, S., Chiang, B., Tindal, G., & Blackburn, M. *Experimental analysis of program components: An approach to research in CSDC's* (Research Report No. 12). Minneapolis: University of Minnesota. Institute for Research on Learning Disablities, 1979.

Deno, S., & Mirkin, P. *Data-based program modification: A manual*. Minneapolis: Leadership Training Institute/Special Education, University of Minnesota, 1977.

Deutsch, M., Fishman, J.A., Kogan, L., North, R. & Whiteman, M. Guidelines for testing minority group children. *Journal of Social Issues*, 1964, *20*(2), 129-145.

Dickson, C.R. Role of assessment in behavior therapy. In P. McReynolds (Ed.) *Advances in psychological assessment* (Vol.3). San Francisco: Jossey-Bass, 1975.

Dilbard, J.D., Houghton, D.W., & Thomas, D.G. The effects of optometric care on educable mentally retarded children. *Journal of Optometric Vision Therapy*, 1972, *3*, 35-57.

Doll, E.A. *Vineland Social Maturity Scale: Condensed manual of directions* (rev. ed.). Minneapolis: American Guidance Service, 1965.

Doll, E.A. *Preschool attainment record*. Circle Pines, Minn.: American Guidance Services, 1966.

Doyle, I.O. Theory and practice of ability testing in ancient Greece. *Journal of the History of the Behavioral Sciences*, 1974, *10*, 202-212.

Dorsen, N., Bender, P., & Neuborn, B. In B. Emerson, Hafer, & Dorsen (Eds.) Political and civil rights in the United States (Vol. 1) Boston: Little Brown, 1976.

Dubois, P. Testing in ancient China." In A. Anastasi (Ed.) Testing problems in perspective. Princeton, N.J.: Educational Testing Service, 1966.

Dubois, P.H. A history of psychological testing. Boston: Allyn & Bacon, Inc., 1970.

DuBose, R.F., Langley, M.B., & Stagg, U. Assessing severely handicapped children. Focus on exceptional children, 1977, 7, 1-13. (a)

Dubin, J.A., Osburn, H., & Winick, D.M. Speed and practice: Effects on Negro and white test performances. Journal of Applied Psychology, 1959, 53, 19-23.

Dudek, S.Z., Lester, E.P., Goldberg, J.R., & Dyer, G.B. Relationship of Piaget measures to st!ndard intelligence and motor scales. Perceptual and Motor Skills, 1969, 28, 351-362.

Dunn, L. Special education for the mildly retarded: Is much of it justifiable? Exceptional Children, 1968, 35, 5-22.

Dunn, L.M., & Kirk, S.A. Impressions of Soviet psycho-educational service and research in mental retardation. Exceptional Children, 1963, 29, 299-311.

Dunsing, J.D., & Kephart, N.C. Motor generalization in space and time. In J. Hellmuth (Ed.), Learning disorders, Vol. 1, Seattle, Wash.: Special Child Publications, 1965.

Dusek, J.B. The development of test anxiety in children. In I.G. Sarason (Ed.) *Test anxiety: Theory, research, and applications*. Hillsdale, N.J.: Erlbaum Associates, 1980.

Dyer, P.J. *Effects of test conditions on Negro-white differences in test scores*. Doctoral dissertation, Columbia University, 1970.

Early, G.H. & Sharpe, T.M. Perceptual-motor training and basic abilities. *Academic Therapy*, 1970, 5, 235-240.

Ebel, R.L. Some limitations of criterion-referenced measurement. In *Testing in turmoil: A conference on problems in issues in educational measurement*. Greenwich, Conn.: Educational Records Bureau, 1970.

Eckberg, D.L. *Intelligence and race: Origins and dimensions of the IQ controversy*. New York: Praeger, 1979.

Edlund, C.V. The effect on the test behavior of children, as reflected in the IQ scores when reinforced after each correct response. *Journal of Applied Behavior Analysis*, 1972, 5, 317-319.

Edwards, A.J. *Individual mental testing. Part I History and Theories*. Scranton, PA.: Intext Educational Publishers, 1971.

Eells, K. et al. *Intelligence and cultural differences*. Chicago: University of Chicago Press, 1951.

Eells, K., Davis, R., Havinghurst, V., & Tyler, R.W. *Intelligence and cultural differences*. Chicago: University of Chicago Press, 1951.

Egeland, B. Examiner expectancy: Effects on the scoring o& the WISC. *Psychology in the Schools*. 1961, *6*, 313-315.

Eibl-Eibesfeldt, I. *Love And Hate*. New York: Holt, Rinehart and Winston, 1971.

Eisenberg, L., Berlin, C., Dill, A., & Sheldon, F. Class and race effects on the intelligibility of monosyllables. *Child Development*, 1968, *39*, 1077-1089.

Einhorn, H.J., & Bass, A.R. Methodological consideratons relevant to discrimination in employment testing. *Psychological Bulletin*, 1971, *75*, 261-269.

Ekman, P. (Ed.). *Darwin and facial expression*. New York: Academic Press, 1973.

Elashoff, J.D., & Snow, R.E. *Pygmalion reconsidered: A case study in statistical inference: Reconsideration of the Rosenthal-Jacobson data on teacher expectancy*. Worthington, Ohio: Jones, 1971.

Elkind, D. *Children and adolescents: Interpretive essays on Jean Piaget* (2nd ed.). New York: Oxford University Press, 1974.

Ellis, A. *Reason and emotion in psychotherapy*. New York: Lyle Stuart, 1962. (b)

Emery, R.E., & Marholin, D. An applied behavior analysis of delinquency: The relevancy of relevant behavior. *American Psychologist*, 1977, *32*, 860-873.

Epps, E.G. *Situational Effects* in testing. In L.P. Miller (Ed.), *The testing of black students*. Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Epps, E.G., Katz, I., Perry, A., & Runyon, E. Effects of race of comparison reference and motives on Negro cognitive performance. Journal of Educational Psychology, 1971, 62 (3), 201-208.

Evans, I.M. & Nelson, R.O. Assessment of child behavior problems. In A.R. Ciminero, K.W. Calbourn & H.E. Adams (Eds.) Handbook of behavior assessment. New York: Wiley, 1977.

Eysenck, H.J. Behavior therapy and the neuroses. Oxford: Pergamon Press, 1960.

Eysenck, H.J. (Ed.), Experiments in behavior therapy. Oxford: Pergamon Press, 1964.

Farber, B. Mental retardation: Its social context and social consequences. Boston: Houghton Mifflin Company, 1968.

Feagans, L. Ecological theory as a model for constructing a theory of emotional disturbance. In W.C. Rhodes and M.L. Tracy (Eds.) A study of child variance Volume I: Conceptual models. Ann Arbor: University of Michigan, 1972.

Fear, R.A. The evaluation interview. (2nd Ed.) New York: McGraw-Hill, 1973.

Ferster, C. , & Skinner, B.F. Schedules of reinforcement. New York: Appleton-Century-Crofts, 1957.

Feuerstein, R. A dynamic approach to the causation, prevention, and alleviation of retarded performance. In H.C. Haywood (Ed.), Social-cultural aspects of mental retardation. New York: Appleton-Century-Crofts, 1970.

Feuerstein, R. Learning potential assessment device. In B.W. Richards (Ed.), Proceedings of the first conference of the International Association for the Scientific Study of Mental Deficiency. Reigate Surry, England: Jackson, 1968.

Feuerstein, R., Miller, R., Rand, Y., Jensen, J.R. Can evolving techniques better measure cognitive change? The Journal of Special Education, 1981, 15, 201-219.

Feuerstein, R., & Rand, R. The dynamic assessment of retarded performers: The learning potential assessment, device, theory, instruments, and techniques. Baltimore: University Park Press, 1978.

Feuerstein, R., Rand, Y., Hoffman, M.B., & Miller, R. Instrumental enrichment. Baltimore: University Park Press, 1980.

Feuerstein, R., Rand, Y., & Hoffman, M.B. The dynamic assessment of retarded performers. Baltimore: University Park Press, 1979.

Figueroa, R.A. The system of multicultural pluralistic assessment. School Psychology Digest, 1979, 8, 28-36.

Figureli, J.K., & Keller, H.R. The effects of training and socioeconomic class upon the accquisition of conservation concepts, Child Development, 1972, 43, 293-298.

Findley, W.G., & Bryan, M.M. Ability grouping: 1970: Status, impact, and alternatives. Athens, GA.: University of Georgia, Center for Educational Improvement, 1971.

Fisher, A. *Four approaches to classification of mental retardation*. Paper presented at the annual meeting of the American Psychological Association, Toronto, August, 1978.

Fiske, D.W. Two worlds of psychological phenomena. *American Psychologist*, 1979, *34*, 733-739.

Flanders, N.A. *Interaction analysis in the classroom: A manual for observers* (Rev. Ed.). Ann Arbor, Mich.: University of Michigan, 1966.

Flaugher, R. The many definitions of test bias. *American Psychologist*, 1978, *33*, 671-679.

Flavell, J. *The Development Psychology of Jean Piaget*. New York: Van Nostrand, 1963.

Forness, S.E. The reinforcement hierarchy. *Psychology in the Schools*, 1973, *10*, 168-177.

Forrest, E.B. Aproaching vision training. *Academic Therapy*, 1968, *3*, 155-161.

Foster, R. *Camelot behavioral checklist*. Parsons, Kans.: Camelot Behavioral Systems, 1974.

Foster, S.L., & Cone, J.D. Current issues in direct observation. *Behavioral Assessment*, 1980, *2*, 313-338.

Foster, G.G., Ysseldyke, J.E., & Reese, J.H. I wouldn't have seen it if I hadn't believed it. *Exceptional Children*, 1975, *41*, 469-473.

Frame, R. *Diagnoses related to school achievement, client's race, and socio-economic status*. Paper presented at the annual meeting of the American Psychological Association. New York, September, 1979.

Friedman, P.R. Legal regulation of applied behavior analyses in mental institutions and prisons. Arizona Law Review, 1975, 17, 39-104.

Frostig, M. Visual perception, integrative functioning and academic learning. Journal of Learning Disabilities. 1972, 5, 1-15.

Frostig, M. Testing as a basis for educational therapy. Journal of Special Education, 1967, 2, 15-34.

Frostig, M. Frostig developmental test of visual perception (3rd ed.), Palo Alto, Ca.: Consulting Psychologists Press, 1961.

Frostig, M., & Horne, D. The Frostig program for the development of visual perception. Chicago: Follett, 1964.

Frostig, M., Lefever, D.W., & Whittlesey, J.R. A developmental test of visual perception for evaluating normal and neurologically handicapped children. Perceptual and Motor Skills, 1961, 12, 383-394.

Fox, W.L., Egner, A.N., Paolucci, P.E., Perlman, P.F., & McKenzie, H.S. An introduction to a regular classroom approach to special education. In E. Deno (Ed), Instructional alternatives for exceptional children. Arlington, Va.: Council for Exceptional Children, 1973.

Fruness, S.C. A comparison of management groups involving the use of the standrad behavior chart and setting performance aims. Unpublished doctoral dissertation, University of Houston, 1973.

Galton, H. Hereditary Genius: An inquiry into its laws and consequences. London: Macmillan, 1869.

Gallagher, B.G. (Ed.) *NAACP report on minority testing*. National Association for the Advancement of Colored People, May 1976.

Galvan, R.R. *Bilingualism as it relates to intelligence test scores and school achievement among culturally deprived Spanish-American children*. Doctoral dissertation, East Texas State University, Ann Arbor, Mich.: University Microfilms, 1967, No. 68-1131.

Garcia, J. The logic and limits of mental aptitude testing. *American Psychologist*, 1981, *36*, 72-1180.

Garcia, R.L. *Unique characteristics of exceptional bilingual students*. Paper presented at the regional meeting of MSOE and BMC, Kansas City, Mo., June 9, 1976.

Garcia, J. IQ: The Conspiracy. *Psychology Today*, 1972, *40*.

Gardner, R.A. On box score methodology as illustrated by three reviews of overtraining reversal effects. *Psychological Bulletin*, 1966, *66*, 416-418.

Gardner, W.M. *Behavior modification in mental retardation*. Chicago: Aldine-Atherton, 1971.

Garfinkle, A.S. *Development of a Battery of Piagetian Logico-Mathematical Concepts*. Master's thesis, University of Colorado, 1975.

Garth, T.R. The problem of racial psychology. *Journal of Abnormal and Social Psychology*, 1922, *17*, 215-219.

Gay, G., & Abrahams, R. Does the pot melt, boil, or brew? Black children and white assessment procedures. *Journal of School Psychology*, 1973, *11*. 330-340.

Gelfand, D.M., & Hartmann, D.P. *Child behavior analysis and therapy*. New York: Pergamon Press, 1975.

Getman, G.N. The mileposts to maturity. *Optometric Weekly*, 1972, *63*, 321-331.

Getman, G.N. *How to develop your child's intelligence*. (6th ed.) Luverne, Minn.: Announcer Press, 1966. (a)

Getman, G.N. The visuomotor complex in the acquisition of learning skills. In J. Hellman (Ed.), *Learning disabilities*. (Vol. 2). Seattle, Wash.: Special Child Publications, 1966. (b)

Getman, G.N. Pre-school perceptual skills: An aid to first grade achievement. *Optometric Weekly*, 1962, *53*, 1749-1753.

Ghiselli, E.E. *The validity of occupational aptitude tests*. New York: Wiley, 1966.

Ginsburg, H., & Opper, S. *Piaget's theory of intellectual development: An introduction*. Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Glaser, R., & Klaus, A.J. Proficiency measurement: Assessing human performance. In R.M. Gagne (Ed.), *Psychological principles in system development*. New York: Holt, Rinehart, & Winston, 1962.

Gleser, G., Gottshalk, L.A., and Springer, K.H. An Anxiety Scale Applicable to Verbal Samples. *Archives of General Psychiatry*, 1961, *5*, 593-604.

Glynn, E.L., Thomas, J.D., & Shee, S.M. Behavioral self-control of on-task behavior in an elementary classroom. *Journal of Applied Behavior Analysis*, 1973, *6*, 105-113.

Golddiamond, I. Singling out self-administered behavior therapies for professional overview: A comment on Rosen. American Psychologist, 1976, 31, 142-147.

Golddiamond, I. Singling out behavior modification for legal regulation: Some effects on patient care, psychotherapy, and research in general. Arizona Law Review, 1975, 17, 105-126.

Goldfried, M.R. Behavioral assessment in perspective. In J.D. Cone and R.P. Hawkins (Eds.) Behavioral assessment: New directions in clinical psychology. New York: Brunner/Mazel, 1976.

Goldfried, M.R. & Davison, G.C. Clinical behavior therapy. New York: Holt, Rinehart and Winston, 1976.

Goldfried, M.R. & Ingling, J.H. The connotative and symbolic meaning of the Bender-Gestalt. Journal of Projective Techniques and Personality Assessment, 1976, 28, 185-191.

Goldfried, M.R. & Kent, R.M. Tradtional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. Psychological Bulletin, 1972, 77, 409-420.

Goldlfried, M.R. & Lineham, M.M. Basic issues in behavioral assessment. In A.R. Ciminero, K.S. Calhoun, and H.E. Adams (Eds.) Handbook of behavioral assessment. New York: Wiley, 1977.

Goldfried, M.R., & Pomeranz, D.M. Role of assessment in behavior modification. Psychological Reports, 1968, 23, 75-87.

Goldfried, M.R., & Sprafkin, J.N. Behavioral personality assessment. Morristown, N.J.: General Learning Press, 1974.

Goldman, R.D., & Hewitt, B.N. An investigation of test bias for Mexican-American college students, *Journal of Educational Measurement*, 1975, *12*, 187-196.

Goldman,, R.D., & Richards, R. The SAT prediction of grades for Mexican-American versus Anglo-American students of the University of California, Riverside. *Journal of Educational Measurement*, 1974, *11*, 129-135.

Goldschmid, M.L., & Benterl, P.M. The dimensions and measurement of conservation. *Child Development*, 1968, *39*, 787-802.

Goldstein, G. Methodological and theoretical issues in neuropsychological assessment, *Journal of Behavioral Assessment*, 1979, *1*, 23-41.

Goldstein, H., Moss, J.W., & Jordan, L.J. *The efficacy of special class training on the development of mentally retarded children* (Cooperative Research Project 619). Washington, D.C.: U.S. Office of Education, 1965.

Goodman, J.F. Is tissue the issue? A critique of SOMPA'S models and tests. *School Psychology Digest*, 1979, *8*, 47-70.

Goodman, J.F. "Ignorance" versus "stupidity" - the basic disagreement. *School Psychology Digest*, 1979, *8*, 218-223.

Goodman, J.R. The diagnostic fallacy: A critique of Jane Mercer's concept of mental retardation. *Journal of School Psychology*, 1977, *15*, 197-206.

Goodman, L., & Hammill, D. The effectiveness of the Kephart-Getman activities in developing perceptual-motor cognitive skills. *Focus on Exceptional Children*, 1973, *9*, 1-9.

Goodenough, F.L. Mental testing. New York: Rinehart, 1949.

Gordon, J.E., & Haywood, H.C. Input deficits in cultural famiilial retardation. Effects of stimulus enrichment. American Journal of Mental Deficiency, 1969, 73, 604-610.

Gordon, R.A. Examining labeling theory: The case of mental retardation. In W.R. Gove (Ed.), The labeling of deviance: Evaluating a perspective. New York: Halsted Press, 1975.

Gordon, R.L. Interviewing: Strategy, techniques, and tactics (Rev. Ed.). Homewood, Ill, : The Dorney Press, 1975.

Gould, S.J. Ontogeny and phylogeny. Cambridge, Mass: Harvard University Press, 1977.

Gould, S.J. Optometry as a discipline in the educational complex. Optometric Weekly, 1962, 53, 1665-1668.

Gove, W. Societal reaction as an explanation of mental illness: An evaluation. American Sociological Review, 1970, 35 (5), 873-883.

Grabowski, J. & O'Brien, C.P., & Mintz, J. Increasing the likelihood that consent is informed. Journal of Applied Behavior Analysis, 1979, 12, 283-284.

Grant, D.L. , & Gray, D.W. Validation of employment tests for telephone company installation and repair occupations. Journal of Applied Psychology, 1970, 54, 7-14.

Graziano, A.M., DeGiovanni, I.S., & Garcia, K.A. Behavioral treatments of children's fears: A review. Psychological Bulletin, 1979, 86, 804-830.

Green, D.R. Racial and ethnic bias in test construction. Monterey: California Test Bureau, 1972.

Greenspan, S.B. The pediatric optometrist as a coorindator of multidisciplinary care. *Journal of the American Optometric Association*, 1973, *44*, 149-151.

Gridley, G., & Mastenbrook, J. *Research on the need for local norms for the Adaptive Behavior Inventory for Children*. Paper presented at the annual meeting of the American Psychological Association, San Francisco, August, 1977.

Grisso, T., & Vierling, L. Minors' consent to treatment: A developmental perspective. *Professional Psychology*, 1978, *9*, 412-427.

Gross, A.L., & Su, W. Defining a "fair" or "unbiased" selection model: A question of utilities. *Journal of Applied Psychology*, 1975, *60*, 345-351.

Gross, M.L. *The Brain Watchers*. New York: Random House, 1962.

Grossman, H.J. *Manual on terminology and classification in mental retardation*. Washington,DC: American Association on Mental Deficiency, 1977.

Grunder, T.M. Two formulas for dtermining the readability of subject consent forms. *American Psychologist*, 1978, *33*, 773-775.

Guarnaccia, V.J. Factor structure and correlates of adaptive behavior in noninstitutionalized retarded adults. *American Journal of Mental Deficiency*, *80*, 543-547, 1976.

Gump, P.V., Schoogen, P., & Redl, F. The behavior of the same child in different mileus. In R.G. Barker (Ed.), *The stream of behavior: Explorations of its structure content*. New York: Meredith Publishing Co., 1963.

Guskin, S.L. Research on labeling retarded persons: Where do we go from here? (A reaction to MacMillan, Jones, and Aloia). *American Journal of Mental Deficiency*, 1974, 79, 262-264.

Guskin, S.L., Bartel, N.R., & MacMillan, D.L. Perspective of the labeled child. In N. Hobbs (Ed.) *Issues in the classification of children* (Vol.2), San Francisco: Jossey-Bass, 1975.

Haas, H. *The Human Animal*. New York: Dell Publishing Co., 1972.

Hall, E.T. *The Silent Language*. New York: Doubleday, 1959.

Hall, R.V., & Copeland, R.E. *The responsive teaching model: A first step in shaping school personnel as behavioral modification specialists*. Paper presented at the Third Banff International Conference of Behavior Modification, April, 1971.

Hall, V.C., & Turner, R.R. Comparison of imitation and comprehension scores between two lower-class groups and the effects of two warm-up conditions on imitation of the same groups. *Child Development*, 1971, 42, 1735-1750.

Hall, V.C., & Turner, R.R. The validity of the "different language explanation" for poor scholastic performance by black students. *Review of Educational Research*, 1974, 44, 69-81.

Hallahan, D.P., & Kauffman, J.M. *Exceptional children*. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Halliwell, J.W., & Solan, H.A. The effects of a supplemental perceptual program on reading achievement. *Exceptional Children*, 1972, 39, 613-621.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B. Criterion-referenced testing and measurements: A review of technical issues and development. *Review of Educational Research*, 1978, *48*, 1-48.

Hamilton, R.G., & Ingling, J.H. Examiner influence on the Holtzman Inkblot Technique. *Journal of Projective Techniques and Personality Assessment*, 1966, *30*, 553-558.

Hammill, D., Goodman, L. & Wiederholt, J.L. Visual-motor processes: Can we train them? *The Reading Teacher*, 1974, *27*, 469-478.

Haney, W. Validity, vaudeville, and values: A short history of social concerns over standardized testing. *American Psychologist*, 1981, *36*, 1021-1034.

Hardisty, J.H. Mental illness: A legal fiction. *Washington Law Review*, 1973, *48*, 735-762.

Hargadon, F. Access to higher education. *American Psycholgist*, 1981, *36*, 1112-1119.

Haring, N.G. & Krug, D.A. Placement in regular programs: Procedures and results. *Exceptional Children*, 1975, *41*, 413-417.

Haring, N.G., Maddux, L., & Krug, D.A. *Investigation systematic instructional procedures to facilitate academic achievement in mentally retarded disadvantaged children: Final report*. Seattle, Wash.: Washington University Press, 1972.

Harms, L.S. Listener comprehension of speakers of three status groups. *Language and Speech*, 1961, *4*, 109-112.

Harris, S., & Masling, J. Examiner sex, subject sex, and Rorschach productivity. *Journal of Consulting and Clinical Psychology*, 1970, *34*, 60-63.

Hartlage, L.C., & Reynolds, C.R. Neuropsychological Assessment and 4he individualization of instruction. In G.W. Hynd & J.E Obrzut (Eds.) *Neuropsychological assessment and the school-age child*. New York: Grune & Stratton, 1981.

Hartlage, P.L., & Givens, T.S. Common neurological problems of school-age children: Case histories. In C.R. Reynolds and T.B. Gutkin (Eds.), *Handbook of School Psychology*. New York: Wiley & Sons, 1982.

Hartley, D. *Observations on man, his fame, his duty, and his expectations*. Gainesville, Fla.: Scholars Facsimiles and Reprints, 1966 (orginally published 1789).

Hartmann, D.P., Roper, B.L. & Bradford, D.C. Some relationships between behavioral and traditional assessment. *Journal of Behavioral Assessment*, 1979, *1*, 3-21.

Hathaway, S.R., & McKinley, J.C. *The Minnesota Multiphasic Personality Inventory*. Minneapolis: University of Minnesota Press, 1943.

Haynes, S.N. *Principles of behavioral assessment*. New York: Gardner Press, 1978.

Haynes, S.N., & Wilson, C.C. *Behavioral Assessment*. San Francisco: Jossey-Bass, 1979.

Haywood, H.C., & Arbitman-Smith, R. The modification of cognitive functions in slow-learning adolescents. In P. Mittler (Ed.) *Proceedings of the 5th International Congress of the International Association for the Scientific Study of Mental Deficiency*. Baltimore: University Park Press, 1980.

Haywood, H.C., Filler, J.W., Shifman, M.A. & Chatelant, G. Behavioral assessment in mental retardation. In P. McReynolds (Ed.), *Advances in Psychological Assessment III*. San Francisco: Jossey-Bass, 1974.

Haywood, H.C., & Switzky, H.N. Children's verbal abstracting: Effects of enriched input, age, and IQ. *American Journal of Mental Deficiency*, 1974, *78*, 556-565.

Heber, R. A manual on terminology and classification in mental retardation. *American Journal of Mental Deficiency*. Monograph Supplement, 1961.

Heckhaussen, H. *The Anatomy of Achievement Motivation*. New York: Academic Press, 1967.

Heider, F. Thing and medium. *Psychological Issues*, 1959, *11*, 1-35.

Henderson, R.W. Social and emotional needs of culturally diverse children. *Exceptional Children*, 1980, [illegible], 598-605.

Henderson, R.W. & Valencia, R.R. Nondiscriminatory school psychological services beyond nonbiased assessment. In J.R. Bergan (Ed.) *Contemporary school psychology*. Columbus, Ohio: Merrill, in press.

Hersen, M. Historical perspectives in behavioral assessment. In M. Hersen & A.S. Bellack (Eds.) Behavioral assessment: A practical handbook. New York: Pergamon, 1976.

Hersen, M. & Bellack, A.S. Social skills training for chronic psychiatric patients: Rationale, research findings, and future directions. Comprehensive Psychiatry, 1976, 17, 559-580. (b)

Hersen, M. Behavor modification approach to a school-phobia case. Journal of Clinical Psychology, 1970, 26, 128-132.

Hersen, M. Sexual aspects of Rorschach administration. Journal of Projective Techniques and Personality Assessment. 1970, 34, 104-105.

Hersen, M. & Bellack, A.S. (Eds.) Behavior Therapy in the psychiatric setting. Baltimore: The Williams & Wilkins Co., 1978.

Hersen, M., & Greeves, S.T. Rorschach productivity as related to verbal reinforcement. Journal of Personality Assessment, 1971, 35, 436-441.

Hess, E.H. and Polt, J.M. Pupil size as related to interest value of visual stimuli. Science, 1980, 123, 349-350.

Hess, E.H., Seltzer, A.L., and Schlien, J.M. Pupil response of hetero a.d homosexual males to pictures of men and women: A Pilot study. Journal of Abnormal Psychology, 1965, 70, 165-168.

Hilgard, E.R. Introduction to psychology (2nd Ed.). New York: Harcourt, Brace, 1957.

Hobbs, N. *The futures of children.* San Francisco: Jossey-Bass, 1975. (a)

Hobbs, N. (Ed.) *Issues in the classification of children* (Vol. 1) San Francisco: Jossey-Bass, 1975. (b)

Hobbs, N. (Ed.) Issues in the classification of children (Vol. 2) San Francisco: Jossey-Bass, 1975. (c)

Hoffman, B. *The tyranny of testing.* New York: Crowell-Collier Press, 1962.

Hofmeister, A. Integrating criterion-referenced testing and instruction. In W. Hively & M. Reynolds (Eds.), *Domain-referenced testingg in special education.* Minneapolis: Leadership Training Institute/Special Education, University of Minnesota, 1975.

Holland, C.J. An interview guide for behavioral counseling with parent. *Behavior Therapy*, 1970, *1*, 70-79.

Holland, W.R. Language barrier as an education problem of Spanish-speaking children. *Exceptional Children*, 1960, *27*, 42-50.

Holmen, M.G., & Docter, R. *Educational and Psychological Testing*, 1972.

Holtzman, W.H. The changing world of mental measurement and its social significance. *American Psychologist*, 1171, *26*, 546-553.

Hudson, B.A., & Niles, J.A. Trail teaching. The missing link. *Psychology in the Schools*, 1974, *11*, 188-191.

Hugdahl, K. The three-system model of fear and emotion--A critical examination. *Behavior Research and Therapy*, 1981, *19*, 75-85.

Humphreys, L.G. Statistical definitions of test validity for minority groups. *Journal of Applied Psychology*, 1973, *58* (1), 1-4.

Humphreys, L., & Stubbs, J. A longitudinal analysis of teacher expectation, student expectation, and student achievement, *Journal of Educational Measurement*. 1977, *14*, 261-270.

Hunt, J. McV *Intelligence and Experience*. New York: Ronald Press, 1961.

Hunter, C.P. Classroom observation instruments and teacher inservice training by school psychologists. *School Psychhology Monograph*, 1977, *3*(2), 45-88. (b)

Hunter, J.E. *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. Washington, D.C.: U.S. Employment Service, U.S. Department of Labor, 1980.

Hunter, J.E., & Schmidt, F.L. A critical analysis of the statistical and ethical implications of five definitions of test fairness. _Psychological Bulletin_, 1976, _83_, 1053-1071.

Hunter, J.E., & Schmidt, F.L. Differential and single group validity of employment tests by race: A critical analysis of three recent studies. _Journal of Applied Psychology_, 1978, _63_, 1-11.

Hunter, J.E., & Schmidt, F.L., & Hunter, R. Differential validity of employment tests by race: A comprehensive review and analysis. _Psychological Bulletin_, 1979, _86_, 721-735.

Hurlock, E.B. An evaluation of certain incentives used in schoolwork. _Journal of Educational Psychology_, 1925, _16_, 145-159.

Hutt, M.L. _The Hutt adaptation of the Bender-Gestalt test: Revised_. New York: Grune and Stratton, 1968.

Hutt, M.L., & Briskin, G.J. _The clinical use of the revised Bender-Gestalt test_. New York: Grune and Stratton, 1960.

Irons, D. The effect of familiarity with the examiner on WISC-R verbal, performance, and full scale scores. _Psychology in the Schools_, 1981, _18_, 496-499.

Izard, C.E. _The Face of Emotion_. New York: Appleton-Century-Crofts, 1971.

Jackson, C.D. On the report of the Ad Hoc Committee on Educational Uses of Tests with Disadvantaged Students: Another psychological view from the Association of Black Psychologists. _American Psychologist_, 1975, _30_, 86-90

Jacobs, J.G. & DeGraaf, C.A. _Expectancy and race: Their influences upon the scoring of individual intelligence tests_. Final Report, Project No. 1-E-096. Washington, D.C.: U.S. Department of Health, Education, and Welfare, March 14, 1972.

Jacobson, L.I., Berger, S.E., Bergman, R.L., Milham, J., & Greeson, L.E. Effects of age, sex, systematic conceptual learning, acquisition of learning sets, and programmed social interaction on the intellectual and conceptual development of preschool children from poverty backgrounds. _Child Development_, 1974, _45_, 517-521.

Jensen, A.R. How much can we boost I.Q. and scholastic achievement? _Harvard Educational Review_, 1969, _39_, 1-123.

Jensen, A.R. Personality and scholastic achievement in three ethnic groups. _British Journal of Educational Psychology_, 1973, _43_, 115-125.

Jensen, A.R. How biased are culture-loaded tests? _Genetic Psychology Monographs_, 1974, _90_, 185-244.

Jensen, A. Bias in mental testing. New York: The Free Press, 1980.

Jensen, A.R., & Figueroa R.A. Forward and backward digit-span interaction with race and IQ: Predictions from Jensen's theory. Journal of Educational Psychology, 1975, 67, 882-893.

Johnson, O.G. (Ed.). Tests and measurements in child developments: Handbook II. San Francisco: Jossey-Bass, 1976.

Johnson, S.B. Children's fears in the classroom setting. School Psychology Digest, 1979, 8, 382-396.

Johnson, D., & Mykelbust, H.R. Learning disabilities: Educational principles and practices. New York: Grune & Stratton, 1967.

Johnson, S.M., & Bolstad, O.D. Methodological issues in naturalistic observations: Some problems and solutions for field research. In L.A. Hamerlynck, L.C. Handy, & E.J. Mash (Eds.), Behavior change: Methodology, concepts, and practice. Champaign, Ill.: Research Press, 1973.

Jones, R.R. Behavioral observation of frequency data: Problems in scoring. In L.A. Hamerlynck, L.C. Handy, & E.J. Mash (Eds.) Behavior change: Methodology, concepts and practice. Champaign, Ill.: Research Press, 1973

Jones, R.R., Reid, J.B., & Patterson, M.R. Naturalistic observations in clinical assessment. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. 3). San Francisco: Jossey-Bass, 1974.

Kagan, J. Controversies in intelligence: The meaning of intelligence. In D.W. Allen & E. Seifman (Eds.), The Teacher's Handbook, pp. 655-662, Glenview, Ill.: Scott, Foresman, 1971.

Kahn, R.L. & Cannell, C.F. The dynamics of interviewing: Theory techniques, and cases. New York: John Wiley and Sons, 1957.

Kallingal, A. The prediction of grades for black and white students of Michigan State University. Journal of Educational Measurement, 1971, 8, 263-265.

Kallman, W.M. & Feurerstein, M. Psychophysiological procedures. In A.R. Ciminero, K.S. Calhoun & H.E. Adams (Eds.), Handbook of behavioral assessment New York: John Wiley & Sons, 1977.

Kamin, L.J. The science and politics of I.Q. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1974.

Kane, M. Summary report: Experimental program-transitional first grade. Journal of Optometric Vision Therapy, 1972, 3, 23-29.

Kanfer, F.H. Self-management methods. In F.H. Kanfer & A.P. Goldstein (Eds.) 1975.

Kanfer, F.H. & Grimm, L.G. Behavior analysis: Selecting target behavior in the interview. Behavior Modification, 1977, 1, 7-28.

Kanfer, F.H., & Phillips, J.S. Learning foundations of behavior therapy. New York: Wiley, 1970.

Kanfer, F.H., & Saslow, G. Behavioral diagnosis. In C.M. Franks (Ed.) Behavior therapy: Appraisal and status. New York: McGraw-Hill, 1969, 417-444.

Kass, R.E. & O'Leary, K.D. *The effects of observer bias in field-experimental settings.* Paper presented at a symposium entitled "Behavior Analyses in Education". University of Kansas, Lawrence, April,1970.

Karz, I., Atchison, C.O., Epps, E.G., & Roverts, S.). Race of evaluator, race of form, and expectancy as determinants of black performmance. *Journal of Experimental Social Psychology*, 1972, *8*, 1-15.

Katzell, R.A., & Dyer, F.J. Differential validty revived. *Journal of Applied Psychology*, 1977, *62*, 137-145.

Kaufman, A.S. Cerebral specialization and intelligence testing. *Journal of Research and Development in Education*, 1979, *12*, 96-108. (a)

Kaufman, A.S. Piaget and Gesell: A psychometric analysis of tests built from their tasks. *Child Development*, 1971, *42*, 1341-1360.

Kazdin, A.E. Acceptability of alternative treatments for deviant child behavior. *Journal of Applied Behavior Analysis*, 1980, *13*, 259-273 (a).

Kazdin, A.E. *Behavior modification in applied settings.* (Rev. ed.) Homewood, Ill.:Dorsey Press, 1980. (b)

Kazdin, A.E. *Research design in clinical psychology.* New York: Harper & Row, 1980.(c)

Kazdin, A.E. Fictions, factions and functions of behavior therapy. *Behavior Therapy*, 1979, *10*, 629-654. (a)

Kazdin, A.E. Situational specificity: The two-edged sword of behavioral assessment. Behavioral Assessment, 1979, 1, 57-75. (b)

Kazdin, A.E. Unobtrusive measures in behavioral assessment. Journal of Applied Behavior Analysis, 1979, 12, 713-724. (c)

Kazdin, A.E. History of behavior modification: Experimental foundations of contemporary research. Baltimore: University Park Press, 1978.

Kazdin, A.E. Artifact, bias, and complexity of assessment. The ABC's of reliability. Journal of Applied Behavior Analysis, 1977, 10, 141-150.

Kazdin, A.E. The token economy: A review and evaluation. New York: Plenum Press, 1977.

Kazdin, A.E. Assessing the clinical or applied significance of behavior change through social validaton. Behavior Modification, 1977, 1, 427-452. (b)

Kazdin, A.E. Characteristics and trends in applied behavior analysis. Journal of Applied Behavior Analysis, 1975, 8, 332.

Kazdin, A.E. Self-monitoring and behavior change. In M.J. Mahoney & C.E. Thoresen (Eds.), Self-control: Power to the person. Monterey, Calif.: Brooks/Cole, 1974.

Kazdin, A.E., & Hersen, M. The current status of behavior therapy. Behavior Modification, 1980, 4, 283-302.

Kazdin, A.E. & Wilson, G.T. Evaluation of behavior therapy: Issues, evidence, and research strategies, Cambridge, Mass: Ballinger, 1978.

Kazimour, K., & Reschly, D. *The relationship of the ABIC to ability, achievement, and sociocultural background*. Unpublished manuscript, Iowa State University, 1980.

Keliher A.V. *A critical study of homogeneous grouping* (no. 452, Contributions to Education). New York: Columbia University, Teachers College, 1931.

Kelman, H.C. *A time to speak: On human values and social research*. San Francisco: Jossey-Bass, 1968.

Kelman, H.C. The rights of the subject in social research. An analysis in terms of relative power and legitimacy. *American Psychologist*, 1971, *26*, 989-1016.

Kendall, P.C. Assessment and cognitive-behavioral interventions: Purposes, proposals, and problems. In P.C. Kendall and S.D. Hollond (Eds.) *Assessment strategies for cognitive-behavioral interventions*. New York: Academic Press, 1981. (a)

Kendall, P.C. Cognitive-behavioral interventions with children. In B.B. Lahey & A.E. Kazdin (Eds.), *Advances in clincal child psychology*. (Vol. 4) New York, Plenum , 1981. (b)

Kent, R.M., & Foster, S.L. Direct observational procedures: Methodological issues in appliled settings. In A.R. Ciminero, K.S. Calhoun, & H.E. Adams (Eds.) *Handbook of behavioral assessment*. New York: Wiley, 1977.

Kent, R.M., O'Leary K.D., Diamert, C., & Dietz, S. Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology*, 1974, *42*, 774-780.

Keogh, B.K. Optometric vision training programs for children with learning disabilities: Review of issues and research. Journal of Learning Disabilities, 1974, 1, 36-48.

Kephart, N.C. The slow learner in the classroom. (2nd ed.) Columbus, Ohio: Charles E. Merrill, 1971.

Kephart, N.C. Perceptual motor aspects of learning disabilities. Exceptional Children, 1964, 31, 201-206.

Kephart, N.C. The slow learner in the classroom. Columbus, Ohio: Charles E. Merrill, 1960.

Keston, M.J., & Jimenez, C. A study of the performance on English and Spanish editions of the Stanford-Binet Intelligence Test by Spanish American children. Journal of Gentetic Psychology, 1954, 85, 263-269.

Kimura, D. Functional Asymmetry of the human brain in dichotic listening. Cortex, 1967, 3. 153-178.

Kirk, S.A., & Kirk, W.D. Psycholinguistic learning disabilities: Diagnosis and remediation. Urbana, Ill: University of Illinois Press, 1971.

Kirk, S.A., McCarthy, J.J., & Kirk, W.D. Illinois Test of Psycholinguistic Abilities. Urbana, Ill.: University of Illinois Press, 1971.

Kirk, S.A., McCarthy, J.J., & Kirk, W.D. Examiner's Manual: Illinois test of psycholinguistic abilities. Urbana, Ill: University of Illinois Press, 1968.

Kirkland, K.D. & Thelen, M.H. Uses of modeling in child treatment. In B.B. Lahey and A.E. Kazdin (eds.), Advances in Clinical Child Psychology. Vol. 1. New York: Plenum, 1977.

Klausmeier, H.J., and Goodwin, W. *Learning and human abilities: Educational psychology*. New York: Harper & Row, 1971.

Klineberg, O. *Race differences*. New York: Harper, 1935.

Klineberg, O. Tests of Negro intelligence. In O. Klinebert (Ed.), *Characteristics of the American Negro*. New York: Harper, 1944.

Klugman, S.F. The effects of money incentive versus praise upon the reliability and obtained scores of the Revised Stannford-Binet Test. *Journal of Genetic Psychology*, 1944, 30, 255-269.

Knox H.A. A scale based on the work at Ellis Island for estimating mental defect. *Journal of the American Medical Association*, 1914, *62*, 741-747.

Kolhberg, L. Early education: A cognitive-development view. *Child Development*, 1968, *39*, 1013-1062.

Koocher, G.P. A bill of rights for children in psychotheropy. G.P. Koochen (Ed.) *Children's rights and the mental health professions*. New York: Wiley, 1976.

Korchin, S.J., & Schuldberg, D. The future of clinical assessment. *American Psychologist*, 1981, *36*, 1147-1158.

Kraepelin, E. *One Hundred Years of Psychiatry*. (W. Baskin, trans.) New York: Citadel, 1962.

Krasner, L. Behavior modification: Ethical issues and future trends. In H. Letenverg (Ed.) *Handbook of behavior modification and behavior therapy*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.

Kratochwill, T.R. The movement of psychological extras into ability assessment. Journal of Special Education, 1977, 11, 299-311.

Kratochwill, T.R. N=1: An alternative research strategy for school psychologists. Journal of School Psychology, 1977, 15, 239-249.

Kratochwill, T.R. (Ed.) Single subject research: Strategies for evaluating change. New York: Academic Press, 1978.

Kratochwill, T.R. Behavioral assessment of academic and social problems: Implications for the individual education program. School Psychology Review, 1980, 9, 199-206.

Kratochwill, T.R. Advances in behavioral assessment. In C.R. Reynolds and T.B. Gutkin (Eds.) Handbook of school psychology. New York: Wiley, 1982.

Kratochwill, T.R. Child behavioral assessment: Issues, developments, and directions. Manuscript submitted for publication, 1983.

Kratochwill, T.R., Alper, S., & Cancelli, A.A. Nondiscriminatory assessment in psychology and education. In L. Mann & D.A. Sabatino (Eds.), Fourth review of special education. New York: Grune & Stratton, 1980.

Kratochwill, T.R. & Bergan, J.R. Evaluating programs in applied settings through behavioral consultaton. Journal of School Psychology, 1978, 16, 375-386. (a)

Kratochwill, T.R. & Bergan, J.R. Training school psychologists: Some perspectives on a competency-based behavioral consultation model. Professional Psychology, 1978, 9, 71-82. (b)

Kratochwill, T.R., & Piersel, W.C. Time-series research: Contributions to empirical clinical practice. Behavioral Assessment in press.

Kratochwill, T.R., & Severson, R.A. Process assessment: An examination of reinforcer effectiveness and predictive validity. Journal of School Psychology, 1977, 15, 293-300.

Krauss, R.M., & Rotter, G.S. Communication abilities of children as a function of status and age. Merill-Palmer Quarterly, 1968, 14, 161-173.

Kroeber, A.L. & Kluckhohn, C. Culture: A critical review of concepts and definitions. New York: Vintage Books, 1952.

Labov, W. The logic of nonstandard English. In F. Williams (Ed.), Language and poverty. Chicago: Markham, 1970.

Lambert, N. AAMD adaptive behavior scale school edition. Diagnostic and technical manual. Monterey, California: Publishers Test Service, 1981.

Lambert, N.M., Windmiller, M., & Cole, L.J. AAMD adaptive behavior scale, public school version. Washington, DC: American Association on Mental Deficiency, 1975.

Lambert, N.M., & Nicoll, R.C. Dimensions of adaptive behavior of retarded and non-retarded public-school children, American Journal of Mental Deficiency, 81, 135-146, 1976.

Lang, P.J. A bio-informational theory of emotional imagery. Psychophysiology, 1979, 16, 495-512.

Lang, P.J. Physiological assessment of anxiety and fear. In Behavioral Assessment: New directions in clinical psychology, (Ed.) J.D. Cone and R.P. Hawkins. New York: Brunner/Mazel, 1977.

Lang, P.J. The application of psychophysiological methods to the study of psychotherapy and behavior modification. In A.E. Bergin & S.L. Garfield (Eds.), Handbook of psychotherapy and behavior change. New York: Wiley, 1971.

Lang, P.J. Fear reduction and fear behavior: Problems in treating a construct. In Research in Psychotherapy, Vol. 3 (Ed.) J.M. Shlien. Washington, D.C.: American Psychological Association, 1968.

Langer, E.J., & Abelson, R.P. A patient by any other name...: Clinician group difference in labeling bias. Journal of Consulting and Clinical Psychology, 1974, 42, 4-9.

Laosa, L.M. Nonbiased assessment of children's abilities: Historical antecedents and current issues. In T. Oakland (Ed.), Psychological and educational assessment of minority children. New York: Brunner/Mazel, 1978.

Laosa, L.M. Nonbiased assessment of children's abilities: Historical antecedents and current issues. In T. Oakland (Ed.) Psychological and educational assessment of minority children. New York: Brunner/Mazel, 1977. (a)

Laosa, L.M. Socialization, education, and continuity: The importance of the sociocultural context. Young Children, 1977. (b)

Laosa, L.M. Cross-cultural and subcultural research in psychology and education. Interamerican Journal of Psychology (Revista Interamericana de Psicologia), 1973, 7, 241-248. (a)

Laosa, L.M. Reform in educational and psychological assessment: Cultural and linguistic issues. Journal of the Association of Mexican American Educators, 1973, 1, 19-24. (b)

Laosa, L.M. & Oakland, T.E. Social control in mental health: Psychological assessment and the schools. Paper presented at the 51st Annual Meeting of the American Orthopsychiatric Association,, San Francisco, April 1974.

Lawler, J.M. I.Q., Heritability and Racism. New York: International Publishers, 1978.

Lazarus, A.A. Multimodal behavior therapy. New York: Springer, 1976.

Lemert, E. Paranoia and the dynamics of exclusion. Sociometry, 1962, 25, 2-20.

Lerner, B. The Supreme Court and the APA, AERA, NCME, test standards: Past references and future possibilities. American Psychologist. 1978, 33, 915-919.

Lerner, J.S. Children with learning disabilities. Boston: Houghton-Mifflin, 1971.

Lerner, J.S. Children with learning disabilities: Theories, diagnosis, teaching strategies (2nd. Ed.) Boston: Houghton Mifflin, 1976.

Levin, J.R. *Learner differences: Diagnosis and prescription.* Hinsdale, Ill.: Dryden Press, 1977.

Levine, S., Elzey, F.F., & Lewis, M. *California preschool social competency scale (CPSCS).* Palo Alto, Ca.: Consulting Psychologists Press, 1969.

Lewin, K. Psychological ecology. In Cartwright, D. (Ed.). *Field theory in social science: Selected theoretical papers by Kurt Lewin.* New York: Harper & Row, 1951.

Liberty, K. *Decide for progress: Dynamic aims and data decisions.* Seattle: Experimental Education Unit, Child Development and Mental Retardation Center, University of Washington, 1975.

Liddle, M.P. The schools psychologist's role with the culturally handicapped. In J.F. Magary (Ed.), *School psychological services in theory and practice.* Englewood Cliffs, N.J.: Prentice-Hall, 1967.

Lilienthal, R.A., & Pearlman, K. *The validity of federal selection tests for aid/technicians in the health, science, and engineering fields.* Washington, D.C.: U.S. Office of Personnel Management, Personnel Research and Development Center, in press.

Linden, K.W. and Linden, J.D. *Modern Mental Measurement: A Historical Perspective.* Boston: Houghton Mifflin Co., 1968.

Lindsley, O.R. Precision teaching in perspective: An interview with Ogden R. Lindsley. *Teaching Exceptional Children,* 1971, *3*, 114-119.

Lindsley, O.R. Direct measurement and prosthesis of retarded behavior. *Journal of Education*, 1964, *147*, 62-81.

Lineham, M.M. Issues in behavioral interviewing. In J.D. Cone and R.P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazell, 1977.

Linn, R.L. Fair test use in selection. *Review of Educational Research*, 1973, *43*, 139-161.

Linn, R.L., & Werts, E.E. Considerations for studies of test bias. *Journal of Educational Measurement*, 1971, *8*, 1-4.

Lipinski, D.P., & Nelson, R.O. Problems in the use of naturalistic observaton as a means of behavioral assessment. *Behavior Therapy*, 1974, 5, 341-351.

Lipowski, Z.J. Psychosomatic medicine in the seventies: An overview. *American Journal of Psychaitry*, 1977, *134*. 233-244.

Livingston, S.a. Psychometric techniques for criterion-referenced testing and behavioral assessment. In J. D. Cone and R.P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/ Mazel, 1977.

Locatis, C.N., & Gooler, D.O. Evaluating second order consequences: Technology assessment and education. Review of Educational Research, 1975, 45, 327-353.

London, P., & Rosenhan, D. (Eds.). Foundations of abnormal psychology. New York: Holt, 1969.

Lord, F.M. A study of item bias, using item characteristic curve theory. In Y.H. Pooringa (ed.), Basic problems in cross-cultural psychology. Amsterdam: Swets & Zeitlinger, 1977.

Lord, F.M. Applications of item resonse theory to practical testing problems. Hillsdale, N.J.: Erlbaum, 1980.

Lovitt, T.C. Applied behavior analysis and learning disabilities --Part 1: Characteristics of ABA, general recommendations, and methodological limitations. Journal of Learning Disabilities, 1975, 8, 432-443.

Lynch, W.W. Guidelines to the use of classroom observation instruments by school psychologists. School Psychology Monograph, 1977, 3, 1-22.

MacMillan, D.L. *Mental retardation in school and society*. Boston: Little, Brown and Company, 1977.

Macready, G.B., & Merwin, J.C. Homogeneity within item forms in domain-referenced testing. *Educational and Psychological Measurements*, 1973, 33, 351-361.

MacMillan, D.L., Jones, R.L., & Aloia, G.F. The mentally retarded label: A theoretical analysis and review of research. *American Journnal of Mental Deficiency*, 1974, 79, 241-261.

MacMillan, D.L., & Meyers, C.E. Educaitonal labeling of handicapped learners. In D.C. Berliner (Ed.) *Review of research in education*, (Vol. 7). Washington, D.C.: American Educational Research Association, 1979.

Magdol, M.S. *Perceptual training in the kindergarten*. San Rafael, Ca.: Academic Therapy Publications, 1971.

Mahl, G.F. Disturbances and silences in the patient's speech in psychotherapy. *Journal of Abnormal and Social Psychology*, 1956, 52, 1-15.

Mahoney, M.J. Some applied issues in self-monitoring. In J.D. Cone and R.P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.

Mahoney, M.J. *Cognition and Behavior Modification*. Cambridge, Mass.: Ballinger, 1974.

Mahoney, M.J., & Arnkoff, D. Cognitive and self-control therapies. In S.L. Garfield & A.E. Bergin (Eds.) *Handbook of psychotherapy and behavior change: An empirical analysis*. (2nd Ed.) New York: Wiley, 1978.

Mahoney, M.J., & Ward, M.P. *Psychological assessment: A conceptual approach*. New York: Oxford University Press, 1976.

Maller, J.B., & Zubin, J. The effects of motivation upon intelligence test scores. *Journal of Genetic Psychology*, 1932, *41*, 136-151.

Maloney, D.M., Harper, T.M., Braukmann, C.J., Fixsen, D.L., Phillips, E.L., & Wolf, M.M. Teaching conversation-related skills to pre-deliquent girls. *Journal of Applied Behavior Analylsis*, 1976, *9*, 371.

Maloney, K.B., & Hopkins, B.L. The modification of sentence structure and its relationship to subjective judgments of creativity in writing. *Journal of Applied Behavior Analysis*, 1973, *6*, 425-433.

Maloney, M.P., & Ward, M.P.. *Psychological assessment: A conceptual approach*. New York: Oxford University Press, 1976.

Mann, L. *On the trail of process*. New York: Grune & Stratton, 1979.

Mann, L., Proger, B., & Cross, L. *Aptitude-treatment interactions with handicapped children: A focus on the measurement of the aptitude component*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April, 1973 (ERIC document 075-520)

Marholin, D. II, & Bijou, S.W. Behavioral assessment: Listen when the data speak. In D. Marholin II (ED.), *Child behavior therapy*. New York: Gardner Press, 1978.

Marlatt, G.A., Demming, B., & Reid, J.V. Loss of control drinking of alcoholics: An experimental analogue. *Journal of Abnormal Psychology*, 1973, *81*, 233-241.

Martin, R. *Legal challenges to behavior modification: Trends in schools, corrections and mental health*. Champaign, Ill: Research Press, 1975.

Martin, R. *Educating handicapped children: The legal mandate*. Champaign, Ill: Research Press, 1979.

Martin, R., & Harmon, D.B. *A preliminary report: Winter Haven study of perceptual learning*. Winter Haven, Fla.: Winter Haven Lions Research Foundation, 1962.

Martinez, O.G. Foreword. *Bilingual testing and assessment, proceedings of BABEL workshop and preliminary findings: Multilingual assessment program*. Berkeley, CA: Bay Area Bilingual Education League, 1972.

Marwit, S.J. Communication of tester bias by means of moedling. *Journal of Projective Techniques and Personality Assessment*, 1969, *33*, 345-352.

Marwit, S.J. , & Marcia, J. Tester bias and response to projective instruments. *Journal of Consulting Psychology*, 1967, *37*, 253-258.

Mash, E. J. & Terdal, L. J. (Eds.). *Behavior therapy assessment: Diagnosis, design, and evaluation*. New York: Springer Publishing Co., 1976.

Mash, E. J. & Terdal, L. G. (Eds.). *Behavioral assessment of childhood disorders*. New York: The Guilford Press, 1981. (a)

Mash, E. J. & Terdal, L. G. Behavioral assessment of childhood disturbance. In E. J. Mash & L. G. Terdal (Eds.), *Behavioral asesssment of childhood disorders*. New York: The Guilford Press, 191. (b)

Masling, J. The influence of situational and interpersonal variables in projective testing. *Psychological Bulletin*, 1960, *57*, 65-85.

Masling, J. Differential indoctrination of examiners and Rorschach response. *Journal of Consulting Psychology*, 1965, *29*, 198-201.

Mastenbrook, J. *Analysis of the concept of adaptive behavior and two assessment instruments* Paper presented at the annual meeting of the American Psychological Association. San Francisco, August, 1977.

Matarazzo, J. E. & Wiens, A. M. Black Intelligence Test of Cultural Homogeneity and Weschler Adult Intelligence Scale scores of black and white police applicants. *Journal of Applied Psychology*, 1974, *62*, 57-63.

Matarazzo, J.D., & Wiens, A.N. The interview: Research on its anatomy and structure. Chicago: Aldine, 1972.

MAI-SEA-CAL Oral Proficiency Tests. Seattle: Seattle Public Schools, 815 Fourth Avenue, N. Seattale, Wash, 98109.

Matuszek, P. & Oakland, T. Factors influencing teachers' and psychologists' recommendations regarding special class placement. Journal of School Psychology, 1979, 17, 116-125.

Matza, D. Becoming deviant. Englewood Cliffs, N.J.: Prentice-Hall, 1969.

May Lan, S.P., Lowenstein, R., Sinnette, C., Rogers, C., & Novick, L. Screening and referral outcomes of school-based health services in a low-income neighborhood. Public Health Reports, 1976, 91, 514-520.

Mead, G. H. Self and society. Chicago: The University of Chicago Press, 1934.

Medley, D. M. & Metzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand MacNally, 1963.

Meehl, P. Wanted--a good cookbook. American Psychologist, 1956, 2, 263-272.

Meichenbaum, D. Cognitive-behavior modification: An integrative approach. New York: Plenum Press, 1977.

Meichenbaum, D.H. Cognitive behavior modification. Morristown, N.J.: General Learning Press, 1974.

Melamed, B. B. & Siegel, L. J. Behavioral medicine: Practical applications in health care. New York: Springer, 1980.

Melton, G. B. Children's participation in treatment planning: Psychological and legal issues. *Professional Psychology*, 1981, 12, 246-252.

Mercer, J. R. Sociological perspectives on mild mental retardation. In H.C. Haywood (Ed.). *Social-cultural aspects of mental retardation*, New York: Appleton-Century-Crofts, 1970.

Mercer, J. R. Sociocultural factors in labeling mental retardates. *Peabody Journal of Education*, 1971, *48*, 188-203.

Mercer, J. R. IQ: The lethal label. *Psychology Today*, September 1972, 44.

Mercer, J. R. Implications of current assessment procedures for Mexican American children. *Journal of the Association of Mexican American Educators*, 1973, *1*, 25-33.

Mercer, J. R. *Labeling the mentally retarded*. Berkeley, CA: University of California Press, 1973b.

Mercer, J. R. Psychological assessment and the rights of children. In N. Hobbs (Ed.) *Issues in the Classification of Children* (Vol. I), San Francisco: Jossey-Bass, 1975.

Mercer, J. R. *Cultural diversity, mental retardation, and assessment: The case for nonlabeling*. Paper presented to the Fourth International Congress of the International Association for the Scientific Study of Mental Retardation, Washington, D.C.: August, 1976.

Mercer, J. In defense of racially and culturally nondiscriminatory assessment. *School Psychology Digest*, 1979, *8*, 89-115(a).

Mercer, J. *Technical manual: SOMPA: System of multicultural pluralistic assessment*. New York: Psychological Corporation, 1979(b).

Mercer, J. & Lewis, J. *Technical manual: SOMPA: System of multicultural assessment*. New York: Psychological Corporation, 1978.

Mercer, J. & Smith, J. M. *Subtest Estimates of the WISC Full Scale IQ's for Children*. Vital and Health Statistics, Series 2, No. 47. Washington, D.C.: Government Printing Office, March 1972.

Mercer, J. R. & Ysseldyke, J. Designing diagnostic intervention programs. In T. Oakland (Ed.), *Psychological and educational assessment of minority childlren*. New York: Brunner/Mazel, 1977.

Merz, W. R. *A factor analysis of the Goodenough-Harris drawing test across four ethnic groups*. (Doctoral dissertation. University of New Mexico). Ann Arbor, MI: University of Micro-films, 1970, 70-19, 714.

Messer, S. B. Reflection-impulsivity: A review. *Psychological Bulletin*, 1976, *83*, 1026-1052.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*, 955-966.

Messick, S. The controversy over coaching: Issues of effectiveness and equity. In B.G. Green (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias*. San Francisco: Jossey-Bass, 1981.

Meyer, V., Liddell, A. & Lyons, M. Behavioral interviews. In A. R. Cimenero, K. S. Calhoun & H. E. Adams (Eds.), *Handbook of behavioral assessment*, New York: Wiley, 1977.

Meyers, P. & Hammill, D. *Methods for learning disorders*. New York: Wiley, 1969.

Meyers, P. & Hammill, D. Deprivation or learning disability: Another dilemma for special education. *Journal of Special Education*, 1973, *7*, 409-411.

Meyers, C., Sundstrom, P. & Yoshida, R. The school psychologist and assessment in special education: A report of the Ad Hoc Committee of APA Division 16. *Monographs of Division 16 of the American Psychological Association*, 1974, *2*, 3-57.

Miele, F. Cultural bias in the WISC. *Intelligence*, 1979, *3*, 149-164.

Miklich, D.R., & Creer, T.L. Self-modeling. In J.C. Cull and R.E. Hardy (Eds.), *Behavior modification in rehabilitation settings: Applied principles*. Springfield, Ill,: Charles C. Thomas, 1974.

Miller, L. P. (Ed.) *The testing of black students: A symposium*. Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Miller, R. & Willner, H. S. The two-part consent form: A suggestion for promoting free and informed consent. *The New England Journal of Medicine*, 1974, *290*, 964-966.

Minkin, N., Braukmann, C.J., Minkin, B.L., Timbers, G.D., Timbers, B.J., Fixsen, D.L., Phillips, E.L., & Wolf, M.M. The social validation and training of conversational skills. *Journal of Applied Behavioral Analysis*, 1976, *9*, 127-139.

Minskoff, E., Wiseman, D.E., & Minskoff, J.G. The MWM program for developing language abilities. Ridgewood, N.J.: Educational Performance Associates, 1972.

Mirkin, P.K. A comparison f the effects of three formative evaluation strategies on reading performance. Unpublished doctoral dissertation, University of Minnesota, Minneapolis, 1978.

Mirkin, P.K., Deno, S.L. Tindal, G., & Kuehnle, K. Toward a more systematic approach to periodic review of programs. (Research Report No. 23.) Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1980.

Mischel, W.L. On the interface of cognition and personality: Beyond the personsituation debate. American Psychologist, 1979, 34, 740-754.

Mischel, W. Cognitive appraisals and transformations in self-control. In B. Winers (Ed.), Cognitive views of human motivation. New York: Academic Press, 1974.

Mischel, W.L. Toward a cognitive social learning reconceptualization of personality. Psychological Review, 1973, 80. 252-283.

Mischel, W. Direct versus indirect personality assessment: Evidence and implications. Journal of Consulting and Clinical Psychology, 1972, 38, 319-324.

Mischel, W. Personality and assessment. New York: John Wiley, 1968.

Morganstern, K. Behavioral interviewing: The initial stages of assessment. In M. Hersen & A. Bellack (Eds.), *Behavioral assessment: A practical handbook*. Elmsford, N.Y.: Pergamon Press, 1976.

Morris, R. J. & Brown, D. K. Legal and ethical issues in behavior modification with mentally retarded persons. In J. Matson and F. Andrasik (Eds.), *Treatment issues and innovations in mental retardation*. New York: Plenum, in press.

Morris, R. J. & Kratochwill, T. R. *Assessing and treating children's fears and phobias: A behavioral approach*. New York: Pergamon, 1983.

Mullins, J.B. A rationale for vision training. *Journal of the American Optometric Association*, 1969, *40*, 139-142.

McCormick, E.J., & Tiffin, J. *Industrial psychology* (6th Ed.) Englewood Cliffs, N.J.: Prentice-Hall, 1974.

McClure, W. E. The status of psychological testing in large city public school systems. *Journal of Applied Psychology*, 1930, *14*, 486-496.

McFall, R. M. Analogue methods in behavioral assessment: Issues and prospects. In J. D. Cone and R. P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977. (a)

McFall, R. M. Parameters of self-monitoring. In R. B. Stuart (Ed.), *Behavioral self-management: Strategies, techniques, and outcomes*. New York: Brunner/Mazel, 1977. (b)

McLean, P. D. The effect of informed concert on the acceptance of random treatment assignment in a clinical population. *Behavior Therapy*, 1980, *11*, 129-133.

McNamara, J. R. & Woods, K. M. Ethical considerations in psychological research: A comparative review. *Behavior Therapy*, 1977, *8*, 703-708.

McNamara, J. R. Socioethnical considerations in behavior therapy research and practice. *Behavior Modification*, 1978, *3*, 3-24.

McNemar, Q. On so-called test bias. *American Psychologist*, 1975, *30*, 848-851.

McNemar, Q. Reply to Bass and Angoff. *American Psychologist*, 1976, *31*, 612-613.

McReynolds, P. Historical antecedents of personality assessment. In P. McReynolds (Ed.) *Advances in psychological assessment* (Vol 3), San Francisco: Jossey-Bass Publishers, 1975.

Nalven, F. B., Hoffman, L. J., & Bierbryer, B. The effects of subjects' age, sex, race, and socioeconomic status on psychologists' estimates of "true IQ" from WISC scores. *Journal of Clinical Psychology*, 1969, *25*, 271-274.

National Association of School Psychologists Principles for Professional Ethics. National Association of School Psychologists, P. O. Box 55, Southfield, MI 48037, 1978.

Nay, W. R. Analogue measures. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.) *Handbook of behavioral assessment*. New York: John Wiley & Sons, 1977.

Nelson, R. O. An expanded scope for behavior modification in school settings. Journal of School Psychology, 1974, 12, 276-287.

Nelson, R. O. & Bowles, P. E. The best of two worlds--Observation with norms. Journal of School Psychology, 1975, 13, 3-9.

Nelson, R. O. Assessment and therapeutic functions of self-monitoring. In M. Hersen , Eisler, and Miller (Eds.) Progress in behavior modification (Vol. 5). New York: Academic, 1977.

Nelson, R. O. Realistic dependent measures for clinical use. Journal of Consulting and Clinical Psychology, 1981, 49, 168-182.

Nelson, R. O. & Hayes, S. C. The nature of behavioral assessment: A commentary. Journal of Applied Behavior Analysis, 1979, 12, 491-500.

Newland, T. E. Assumptions underlying psychological testing. In T. D. Oakland & B. N. Phillips (Eds.), Assessing minority group children. A special issue of the Journal of School Psychology. New York: Behavioral Publications, 1978, 315-322.

Newland, T.E. Tested "intelligence" in children. School Psychology Monograph, 1977, 3, 1-44.

Nichols, P. L. The effects of heredity and environment on intelligence test performance in 4- and 7-year-old white and Negro sibling pairs. Doctoral dissertation, University of Minnesota, 1972.

Nihira, K. Factorial dimensions of adaptive behavior in mentally retarded children and adolescents. *American Journal of Mental Deficiency*, *74*, 130-141, 1969.

Nihira, K. Factorial dimensions of adaptive behavior in adult retardates. *American Journal of Mental Deficiency*, *73*, 868-878, 1969.

Nihira, K., Foster, R., Shellhaas, M., & Leland, H. *AAMD adaptive behavior scale: Manual* (Rev. ed.). Washington, D.C.: American Association on Mental Deficiency, 1974.

Nihira, K., Foster, R., Shellhaas, M., & Leland, H. *AAMD adaptive behavior scale*. Washington, D.C.: American Association on Mental Deficiency, 1969.

Nobel, C. E. Race, reality, and experimental psychology. *Perspectives in biology and medicine*, 1959, *13*, 10-30.

Novick, M. R. Federal guidelines and professional standards. *American Psychologist*, 1981, *36*, 1035-1046.

Oakland, T. (Ed.). *Psychological and educational assessment of minority children*. New York: Brunner/Mazel, 1977.

Oakland, T. Research on the adaptive behavior inventory for children and the estimated learning potential. *School Psychology Digest*, 1979, *8*, 63-70.

Oakland, T., & Goldwater, D. Assessment and interventions for mildly retarded and learning disabled children. In G. Phye & D. Reschly (Eds.), *School psychology: Perspectives and isues*. New York: Academic Press, 1979.

Oakland, T. & Laosa, L. M. Professional, legislative and judicial influences on psychoeducational assessment practices on schools. In T. Oakland (Ed.) *Psychological and educational assessment of minority children*. New York: Brunner/Mazel, 1977.

Oakland, T. & Matuszek, P. Using tests in nondiscriminatory assessment. In T. Oakland (Ed.), *Psychological and educational assessment ogf minority children*. New York:Brunner/Mazel, 1977.

O'Connor, R. D. Relative efficacy of modeling, shaping, and the combined procedures for modification of social withdrawal. *Journal of Abnormal Psychology*, 1972, *79*, 327-334.

Odum, E. P. *Fundamentals of ecology*. Philadelphia: W. B. Saunders, 1953.

O'Leary, K. D., Kent, R. M, & Kantowitz, J. Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behavior Analysis*, 1975, *8*, 48-51.

O'Leary, K. D., Romanczyk, R. G., Kass, R. E., Dietz, A., & Santogrossi, D. *Procedures for classroom observation of teachers and children, 1979*. Available from K. Daniel O'Leary, Psychology Department, SUNY at Stony Brook, Stony Brook, New York, 11794.

Ordiz, C. C. & Ball, M. *The Enchilada Test*. Institute for Personal Effectiveness in Children, 1972.

Otis, A. S. An absolute point scale for the group measure of intelligence. *Journal of Educational Psychology*, 1918, *9*, 238-261.

Ozer, M. N. The use of operant conditioning in the evaluation of children with learning problems. *Clinical Proceedings, Children's Hospital of Washington, D.C.*, 1966, *22*, 235.

Ozer, M. N. The neurological evaluation of school-age children. *Journal of Learning Disabilities*, 1968, *1*, 84.

Ozer, M. N. Involving the teacher in the child evaluation process. *Journal of Learning Disabilities*, 1978, *11*, 422-426.

Ozer, M. N., & Dworkin, N. D. The assessment of children with learning problems: An inservice teacher training program. *Journal of Learning Disabilities*, 1974, 7, 15-20.

Ozer, M. N., & Richardson, H. B. The diagnostic evaluation of children with learning problems: A "process" approach. *Journal of Learning Disabilities*, 1974, *7*, 30-34.

Palmer, A. B. & Wohl, J. Voluntary admission forms: Does the patient know what he's signing? *Hospital and Community Psychiatry*, 1972, *23*, 250-252.

Parton, D. A. & Ross, A. O. Social reinforcement of children's motor behavior. A review. *Psychological Bulletin*, 1965, 64, 65-73.

Patterson, G. R., Ray, R. W., Shaw, D. A., & Cobb, J. A. *A manual for coding family interactions (6th revision)*, 1969. Available from ASIS National Auxiliary Publications Service, in care of CCM Information Servides, Inc., 909 Third Avenue, New York, N.Y 10022. Document #01234

Paul, G.L. Behavior modification research: Design and tactics. In C.M. Franks (Ed.) *Behavior therapy: Appraisal and status.* New York: McGraw-Hill, 1969.

Paul, G. L. & Bernstein, D. A. *Anxiety and clinical problems: Systematic desensitization and related techniques*. Morristown, N.J.: General Learning Press, 1973.

Pearlman, I., Schmidt, F. L., & Hunter, J. E. Validity generalization results for tests used to predict training success and job proficiency in clerical occupations. *Journal of Applied Psychology*, 1980, *65*, 373-406.

Pearson, K. *The life, letters, and labors of Francis Galton*. (3 Vols.). Cambridge: University Press, 1914-1930.

Peisach, E. C. Children's comprehension of teacher and peer speech. *Child Development*, 1965, *36*, 467-480.

Perales, A. M. The audio-lingual approach and the Spanish-speaking student. *Hispania*. 1965, *48*, 99-102.

Peter, L.J. *Individual instruction*. New York: McGraw-Hill, 1972.

Peterson, N. S. *An expected utility model for "optimal" selection*. Iowa Testing Programs Occasional Paper No. 10, 1975.

Peterson, N. S. & Novick, M. R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 1976, *13*, 3-29.

Peterson, J. *Early conceptions and tests of intelligence*. Yonkers-on-Hudson, N.Y.: World Book, 1925.

Peterson, R. E. Predictive validity of a brief test of academic aptitude. *Educational and Psychological Measurement*, 1968, *28*, 441-444.

Petrie, P., Brown, K., Piersel, W. C., Frinfrock, S. R., Schelble, M., LeBlanc, C. P., & Kratochwill, T. R. The school psychologist as behavioral ecologist. *Journal of School Psychology*, 1980, *18*, 222-233.

Pettigrew, T. F. *A profile of the Negro American*. Princeton, N.J.: Van Nostrand, 1964.

Pfeifer, C. M., Jr. & Sedlacek, W. E. The validity of academic predictors for black and white students at a predominantly white university. *Journal of Educational Measurement*, 1971, *8*, 253-261.

Phillips, B. N. *School stress and anxiety: Theory, research and intervention*. New York: Humn Sciences Press, 1978.

Phillips, L., Draguns, J. G., & Bartlett, D. P. Classification of behavior disorders. In N. Hobbs (Ed.), *Issues in the classification of children* (Vol 1) San Francisco: Jossey-Bass, 1975.

Phillips, E.L., Phillips, E.A., Wolf, M.M., & Fixsen, D.L. Achievement place: Development of the elected manager system. *Journal of Applied Behavior Analysis*, 1973, *6*, 541-561.

Phillips, E. L., Phillips, E.A., Fixsen, D. L., & Wolf, M. M. Achievement place: Modification of the behaviors of predelinquent boys with a token economy. *Journal of Applied Behavior Analysis*, 1971, *4*, 45-59.

Piersel, W. C., Brody, G. H., & Kratochwill, T. R. A further examination of motivational influences on disadvantaged children's intelligence test performance. *Child Development*, 1977, *48*, 1142-1145.

Piersel, W. C., Brook, G. H., & Kratochwill, T. R. A further examination of motivational influences on disadvantaged children's intelligence test performance. Child Development, 1977, 48, 1141-1145.

Pines, M. The brain changers. New York: New American Library, 1973.

Pinter, R. Intelligence testing. New York: Holt, Rinehart & Winston, 1931.

Plake, B. & Hoover, H. A methodology for identifying biased achievement test items that removes the confounding in an items by groups interaction due to possible group differences in instructional level. Paper presented at the annual meeting of the American Educational Research Association, Toronto, September, 1979.

Platt, J. S. The effect of the modified Raven's Progressive Matrices learning potential coaching procedure on Raven's posttest scores and their correlation value with productive variables of learning disabilities. Unpublish doctoral dissertation, University of Kansas, 1976.

Pressey, S. L. & Teter, M. F. A comparison of colored and white children by means of a group scale of intelligence. Journal of Applied Psychology, 1919, 3, 277-282.

Proshansky, H. M., Ittelson, W. H., & Rivling, L. G. Environmental psychology: Man and his physical setting. New York: Holt, Rinehart & Winston, 1970.

Prugh, D. G., Engel, M., & Moore, W. C. Emotional disturbance in children. In N. Hobbs (Ed.), *Issues in the classification of children* (Vol 1), San Francisco: Jossey-Bass, 19175.

Pryzwansky, W. B. School psychology training and practice. The APA perspective. In T. R. Kratochwill (Ed.), *Advances in school psychology* (Vol 2). Hillsdale, N.J.: Lawrence Erlbaum, 1982.

Quay, H. C. Classification. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood* (2nd ed.). New York: John Wiley & Sons, 1979.

Quay, H. C. Language dialect, age, and intelligence test performance in disadvantaged black children. *Child Development*, 1974, *45*, 463-468.

Quay, H.C. Special education: Assumptions, techniques, and evaluative criteria. *Exceptional Children*, 1973, *40*, 165-170.

Quay, H. C. Patterns of aggression, withdrawal, and immaturity. In H. C. Quay & J. S. Werry (Eds.) *Psychopathological disorders of childhood*. New York: Wiley, 1972.

Quay, H. C. Negro dialect and Binet performance in severely diadvantaged black four-year-olds. *Child Development*, 1972, *43*, 245-250.

Quay, H. C. Language, dialect, reinforcement, and the intelligence test performance of Negro children. *Child Development*, 1971, 42, 5-15.

Rachman, S. The passing of the two-stage theory of fear and avoidance: Fresh possibilities. *Behavior Research and Therapy*, 1976, *14*, 125-134.

Rachman, S. Introduction to behavior therapy. *Behavior Research and Therapy*, 1963, *1*, 4-15.

Rachman, S. Systematic desensitization. *Psychological Bulletin*, 1967, *67*, 93-103.

Rains, P. M., Kidsuse, J. I., Duster, T., & Firedson, E. The labeling approach to deviance. In N. Hobbs (Ed.), *Issues in the classification of children*. San Francisco: Jossey-Bassd, 1975.

Rankin, P. T. Pupil classification and grouping. *Review of Educational Research*, 1931, *1*, 200-230.

Rappaport, D., Gill, M., & Schafer, R. *Diagnostic psychological testing*. Chicago: Year Book Medical Publishing, 1945.

Raskin, L. T. & Nagle, R. J. Modeling effects on the intelligence test performance of test-anxious children. *Psychology in the Schools*, 1980, *17*, 351-355.

Raven, J. C. *Progressive Matrices: A perceptual test of intelligence*, 1938, Individual Form. London: H. D. Lewis, 1938.

Reitan, R. M. A research program on the psychological effects of brain lesions in human beings. In N. R. Ellis (Ed.) *International review of research in mental retardation* (Vol 1). New York: Academic Press, 1966.

Reitan, R. M. Psychological effects of cerebral lesions in children of early school age. In R. M. Reitan & L. A. Davison (Eds.), *Clinical neuropsychology: Current status and applications*. Washington, D.C.: Winston, 1974.

Reschly, D.J. Assessing mild mental retardation: The influence of adaptive behavior, sociocultural states, and prospects for nonbiased assessment. In C.R. Reynolds and T.B. Gutkin (Eds.) *The handbook of school psychology*. New York: John Wiley, 1982.

Reschly, D. J. SOMPA: A symposium-editorial comment. *School Psychology Digest*, 1979, *8*, 4.

Reschly,D. J. Nonbiased assessment. In G. Phye & D. J. Reschly (Eds.) *School psychology: Perspectives and issues*. New York: Academic Press, 1979.

Reschly, D. J. Psychological testing in educational classification and placement. *American Psychologist*, 1981, *36*, 1094-1102.

Reschly, D. J. & Lamprechit, M. Expectancy effects of labels: Fact or artifact? *Exceptional Children*, in press.

Reschly, D. J. & Sabers, D. L. An examination of bias in predictin MAT scores from WISC-R scores for four ethnic-racial groups. *Journal of Educational Measurement*, 1979, *16*, 1-9.

Resnick, D. History of educational testing. In A. K. Widgor & W. R. Garber (Eds.), *Ability testing: Uses, consequences, and controversies*. Washington, D.C.: National Academy Press, 1982.

Reubhausen, D. M. Brian, O. G. Privacy and behavioral research. *Columbia Law Review*, 1965, 65, 1184-1215.

Reynolds, C. R. *Test bias: In God we trust, all others must have data*. Paper presented as an invited address for the APA Division of Evaluation and Measurement to the annual meeting fo the American Psychological Association, Los Angeles, August, 1981.

Reynolds, C. R. & Brown, R. T. Bias in mental testing: An introduction to the issues. In C. R. Reynolds & R.T. Brown (Eds.), *Perspectives for bias in mental testing*. New York: Plenum Press, in press.

Reynolds, M. C. & Balow, B. Categories and variables in special education. In R. L. Jones and D. L. MacMillan (Eds.), *Special education in transition*. Boston: Allyn & Bacon, 1974.

Reynolds, M. C. & Balow, F. Categories and variables in special education. *Exceptional Children*, 1972, 38, 357-366.

Rhodes, W. C. & Tracy, M. L. *A study of child variance, Volume I: Conceptual models*. Ann Arbor: University of Michigan, 1972.

Rice, A. Rhythmic training and board balancing prepares a child for formal learning. *Nations's Schools*, 1962, 6, 72.

Richmond, B.O., & Kicklighter, R. *Children's adaptive behavior scale*. Atlanta: Humanics Limited, 1980.

Riessman, F. *The culturally deprived child*. New York: Harper, 1962.

Rimm, D. C. & Masters, J. C. *Behavior therapy: Techniques and empirical findings*. New York: Academic Press, Inc., 1974.

Roach, E.G., & Kephart, N.C. *The Purdue perceptual-motor survey*. Columbus, Ohio: Charles E. Merrill, 1966.

Robins, L. *Deviant children grown up*. Baltimore: Williams & Williams, 1966.

Rogers, C. R. & Skinner, B. F. Some issues concerning the control of human behavior: A symposium. *Science*, 1956, *124*, 1057-1066.

Rosenshine, B. & Furst, N. The use of direct observation to study teaching. In R. Travers (Ed.), *Second handbook of research on teaching*. Chicago: Rand McNally, 1973.

Rosenthal, T. L. Modeling therapies. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.) *Progress in behavior modification* (Vol 2). New York: Academic Press, 1976.

Rosenthal, T. L. & Bandura, A. Psychological modeling: Theory and practice. In S. L. Garfield and A. E. Bergan (Eds.), *Handbook of psychotherapy and behavior change* (2nd ed.). New York: Wiley, 1978.

Rosenthal, R. & Jacobsen, L. *Pygmalion in the classroom: Teacher expectations and pupils' intellectual development*. New York: Holt, 1968.

Ross, A. O. *Psychological disorders of children: A behavioral approach to theory, research and therapy*. New York: McGraw-Hill, 1974.

Ross, A. O. Psychological aspects of learning disabilities and reading disorders. New York: McGraw-Hill, 1976.

Ross, A. O. Psychological disorders of children: A behavioral approach to theory, research and therapy. New York: McGraw-Hill, 1980.

Ross, M.B., & Salvia, J. Attractiveness as a biasing factor in teacher judgments. American Journal of Mental Deficiency, 1975, 80, 96-98.

Roth, G. & McManis, D. L. Social reinforcement effects on block design performance of organic and nonorganic retarded adults. Americal Jorunal of Mental Deficiency, 1972, 77, 181-189.

Roth, L. H., Meisel, & Lidz, C. W. Tests of competency to consent to treatment. American Journal of Psychiatry, 1977, 134, 279-284.

Rowitz, L. Sociological perspectives on labeling (A reaction to MacMillan, Jones, and Aloia). American Journal of Mental Deficiency, 1974, 79, 265-267.

Ruch, W. W. A reanalysis of published differential validity studies. Paper presented at meeting of American Psychological Association, Honolulu, Hawaii, September 1972.

Rumenik, D. K., Capasso, D. R., & Hendrick, D. Experimenter sex effects in behavioral research. Psychological Bulletin, 1977, 84, 852-877.

Sackett, G. P. (Ed.). Observing behavior (Vol 1): Theory and applications in mental retardation. Baltimore: University Park Press, 1978. (a)

Sackett, G. P. (Ed.). *Observing behavior (Vol. 2): Data collection and analysis methods*. Baltimore: University Park Press, 1978. (b)

Sacks, E. L. Intelligence scores as a function of experimentally established social relationships between child and examiner. *Journal of Abnormal and Social Psychology*, 1952, *47*, 354-358.

Salvia, J. & Podol, J. Effects of visibility of a prepalatal cleft on the evaluation of speech. *Cleft Palate Journal*, 1976, *13*, 361-366

Salvia, J. & Ysseldyke, J. E. *Assessment in special and remedial education*. Boston: Houghton Mifflin, 1978.

Samuda, R. J. *Psychological testing of American minorities: Issues and consequences*. New York: Dodd, Mead, 1975.

Samuda, R. J. From ethnocentrism to a multi-cultural perspective in educational testing. *Journal of Afro-American Issues*, 1975, *3*(1), 4-18.

Samuel, W. Observed IQ as a function of test atmosphere, tester expectation, and race of tester: A replication for female subjects. *Journal of Educational Psychology*, 1977, *69*, 593-604.

Sandoval, J. The WISC-R and internal evidence of test bias with minority groups. *Journal of Consulting and Clinical Psychology*, 1979, *47*, 919-927.

Sandoval, J. & Millie, M. *Accuracy judgments of WISC-R item difficulty for minority groups*. Paper presented at the annual meeting of the American Psychological Association, New York, September 1979.

Sapp, G.L., Horten, W., McElroy, K., & Ray, P. An analysis of ABIC score patterns of selected Alabama school children. In *Proceedings of the National Association of School Psychologists/California Association of School Psychologists and Psychometrists*. San Diego, April, 1979.

Sarason, I. G. *Test anxiety: Theory, research, and applications*. Hillsdale, N.J.: Erlbaum, 1980.

Sarason, S. B. The unfortunate fate of Alfred Binet and school psychology. *Teachers College Record*, 1976, *77*, 579-592.

Sarason, S. B., Davison, K. S., Lighthall, F. F., Waite, R. R., & Ruebush, B. K. *Anxiety in Elementary School Children*. New York: Wiley, 1960.

Sattler, J.M. *Assessment of children's intelligence and special abilities (2nd ed.)*. Boston: Allyn and Bacon, 1982.

Sattler, J. M. Racial "experimenter effects" in experimentation, testing, interviewing, and psychotherapy. *Psychological Bulletin*, 1970, *73*, 137-160.

Sattler, J. M. Racial experimenter effects. In K. S. Miller & R. M. Dreger (Eds.) *Comparative physiological, psychological and sociological studies of Negroes and whites in the United States*. New York: Seminar Press, 1973.

Sattler, J. M. *Assessment of children's intelligence*. Philadelphia: Saunders, 1974.

Sattler, J. M., Hillix, W. A., & Neher, L. A. Halo effect in examiner scoring of intelligence test responses. *Journal of Consulting and Clinical Psychology*, 1970, *34*, 172-176.

Sattler, J. M. & Kuncik, T. M. Ethnicity, socioeconomic status, and pattern WISC scores as variaables that affect psycholgoists' estimates of "effective intelligence." Journal of Clinical Psychology, 1975, 32, 362-366.

Sattler, J. M. & Theye, F. Procedural, situational, and interpersonal variables in individual intelligence testing. Psychological Bulletin, 1967, 68, 347-360.

Sattler, J.M. & Winget, B. M. Intelligence testing procedures as affected by expectancy and IQ. Journal of Clinical Psychology, 1970, 26, 446-448.

Savage, J. E., Jr. & Bowers, N. D. Tester's influence on children's intellectual performance. Washington, D.C.: U.S. Office of Education, 1972. (ERIC Microfiche No. 064 329)

Schwitzgabel, R. K. A contractual model for the protection of the rights of institutionalized mental patients. American Psychologist, 1975, 30, 815-820.

Schwitzgabel, R. L. & Schwitzgabel, R. K. Law and psychological practice. New York: John Wiley & Sons, 1980.

Scannell, D. P. A positive view of standardized tests. Focus on Exceptional Children, 1978, 10, 1-10.

Scheff, T. Being mentally ill. Chicago: Aldine, 1966.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 58, 5-9.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. Validity generalization results for computer programmers. Journal of Applied Psychology, 1980, 65, 643-661.

Schmidt, F. L. & Hunter, J. E. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 1977, *62*, 529-540.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. Validity generalization results for two jobs in the petroleum industry. *Journal of Applied Psychology*, 1981, *66*, 261-273.

Schmidt, F. L., Hunter, J. E., & Pearlman, J. Task differences and validity of aptitute tests in selection: A red herring. *Journal of Applied Psychology*, 1981, *66*, 166-185.

Schmidt, F. L., Hunter, J. E., & Urry, V. M. Statistical power in criterion-related validity studies. *Journal of Applied Psychololgy*, 1976, *61*, 478-485.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology*, 1980, *33*, 705-724.

Schroeder, C. S., Teplin, S., & Schroeder, S. R. An overview of common medical problems encountered in schools. In C. R. Reynolds & T. B. Gutkin (Eds.) *Handbook of school psychology*. New York: Wiley & Sons, 1982.

Schultz, C. B. & Sherman, R. H. Social class, development, and differences in reinforcer effectiveness. *Review of Educational Research*, 1976, *46*, 25-29.

Sedlak, R.A., & Weener, P. Review of research on the Illinois Test of Psycholinguistic Abilities. In L. Mann and D. Sabatino (Eds.) *The first review of special education*, New York: Grune and Stratton, 1973.

Seitz, V., Abelson, W. D., Levine, E., & Zigler, E. Effects of place of testing on Peabody Picture Vocabulary Test scores of disadvantaged Head Start and non-Head Start children. *Child Development*, 1975, *46*, 481-486.

Severson, R. A. The case for the classroom management consultant. *Experimental Publications System*, 1971, *11* (Ms No. 397-26).

Severson, R. A. Behavior therapy with learning disabled children. In M. B. Rosenberg (Ed.), *Educational Therapy* (Vol 3). Seattle: Straub, 1973.

Sewell, T. E. Intelligence and learning tasks as predictors of scholastic achievement in black and white first grade children. *Journal of School Psychology*, 1979, *17*, 325-332.

Sewell, T. E. Shaping the furture of school psychology: Another perspective. In J. E. Ysseldyke & R. A. Weinberg (Eds.), *The future of psychology in the schools: Proceedings of the Spring Hill Symposium*. *School Psychology Review*, 1981, *10*(2).

Sewell, T. E., & Severson, R. A. Learning potential and intelligence as cognitive predictors of achievement in first grade children. *Journal of Educational Psychology*, 1974, *66*, 948-965.

Shapiro, M. H. Legislating the control of behavior control: Autonomy and the coercive use of organic therapies. *Southern California Law Review*, 1974, *47*, 237-356.

Sharp, S. E. Individual psychology: A study in psychological method. *American Journal of Psychology*, 1898-99, *10*, 329-391.

Shaw, C. & MacKay, H. *Social factors in juvenile delinquency*. Vol II of National Committee on Law Observance and Law Enforcement, Report on the Causes of Crime. Washington, D.C.: U.S. Printing Office, 1931.

Shaw, C., & Mackay, H. *Juvenile delinquency and urban areas*. Chicago: University of Chicago Press, 1942.

Shechtman, A. Psychiatric symptoms observed in normal and disturbed black children. *Journal of Clinical Psychology*, 1971, 27, 445-447.

Shockley, W. Models, mathematics and moral obligation to diagnose the origin of Negro IQ deficits. *Review of Educational Research*, 1971, 41, 369-377.

Siegel, L. J. Psychomatic and psychophysiological disorders. In R. J. Morris & T. R. Kratochwill (Eds.), *The practice of child therapy: A testbook of methods*. New York: Pergamon, 1983.

Sidman, M. *Tactics of scientific research*. New York: Basic Books, 1960.

Sigel, I. W. & Olmsted, P. Modifications of cognitive skills among lower-class black children. In J. Hellmuth (Ed.), *Disadvantaged child*, Vol 3, New York: Brunner/Mazel, 1970.

Silverman, R. J., Boa, J. K., & Russel, R. H. *Oral language tests for bilingual students. An evaluation of language dominance and proficiency instruments* (No. 300-75-0329). U. S. Dept of Health, Education, and Welfare, Office of Education, 1976.

Simkins, L. Examiner reinforcement and situational variables in a projective testing situation. *Journal of Consulting Psychology*, 1960, *24*, 541-547.

Simon, B. *Intelligence, psychology, and education: A Marxist critique*. London: Lawrence & Wishart, 1971.

Simon, W. E. Expectancy effects in the scoring of vocabulary items: A study of scorer bias. *Journal of Educational Measurement*, 1969, *6*, 159-164.

Sisk, D. Educational planning for gifted and talented. In J. Y. Kauffman & D.P. Haliahan (Eds.), *Handbook of special education*. Englewood Cliffs, N.J.: Prentice-Hall, 1981.

Sitko, M. C., Fink, A. H., & Gellespie, R. H. Utilizing systematic observation for decision making in school psychology. *School Psychology Monograph*, 1971, *3*, 23-44.

Skinner, B. F. The operational analysis of psychological terms. *Psychological Review*, 1945.

Skinner, B. F. *Science and human behavior*. New York: Free Press, 1953.

Skinner, B. F. *Verbal behavior*. New York: Appleton-Century-Crofts, 1957.

Skinner, B.F. *About behaviorism*. New York: Knopf, 1974.

Skinner, B. F. *Contingencies of reinforcement: A theoretical analyses*. New York: Appleton-Century-Crofts, 1969.

Skinner, B. F. *About behaviorism*. New York: Knopf, 1974.

Sloat, R.S. Optometry: What is it worth to education. *Optometric weekly*, 1971, *62*, 48-51.

Smeets, P. M. & Striefel, S. The effects of different reinforcement conditions on the test performance of multihandicapped deaf children. *Journal of Applied Behavioral Research*, 1975, *8*, 83-89.

Smith, D. Unfinished business with informed consent procedures. *American Psychologist*, 1981, *36*, 22-26.

Smith, D. Trends in counseling and psychotherapy. *American Psychologist*, 1972, *37*, 802-809.

Smith, H.M. Motor activity and perceptual developmennt. *Journal of Health, Physical Education, and Recreation*, 1968, *39*, 28-33.

Smith, M. L. & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 1977, *32*, 752-760.

Snow, R. E. Review of R. Rosenthal and L. Jacobson "Pygmalion in the classroom." *Contemporary Psychology*, 1969, *14*, 197-199.

Solkoff, N. Race of experimenter as a variable in research with children. *Developmental Psychology*, 1972, *7*, 70-75.

Spearman, C. *The abilities of man*. New York: Macmillan, 1927.

Speece, R. G. Conditioning and other technologies used to "treat?", "rehabilitate?", "demolish?" prisoners and mental patients. *Southern California Law Review*, 1972, *45*, 616-681.

Spivack, G. & Swift, M. Classroom behavior of children: A critical review of teacher-administered rating scales. *Journal of Special Education*, 1973, *1*, 55-89.

Staats, A. W. *Social behaviorism*. Homewood, IL.: Sorsey Press, 1975.

Staats, A.W. *Child learning, intelligence, and personality*. New York: Harper and Row, 1971.

Staats, A.W. Behavior analysis and token reinforcement in educational behavior modification and curriculum research. In C.E. Thoreson (Ed.) *Behavior modification in education*. Chicago: University of Chicago Press, 1973.

Standards for Educational and Psychological Tests. Washington, D.C.: American Psychological Association, 1974.

Standards for providers of psychological services. *American Psychologist*, 1977, *32*, 495-505.

Stanley, J. C. & Porter, A. C. Correlation of Scholastic Aptitude Test scores with college grades for Negroes versus whites. *Journal of Educational Measurement*, 1967, 4, 199-218.

Stern, W. *Uber Psychologies der individuellen differenzen* (Ideen zur einer "Differenzellen Psychologie:). Leipzig: Barth, 1900.

Stephens, T.M. *Directive teaching of children with learning and behavioral handicaps*. (2nd Ed) Columbus, Ohio: Chalres E. Merrill, 1976.

Stevens-Long, J. The effect of behavioral context on some aspects of adult disciplinary practice and affect. *Child Development*, 1973, *44*, 476-484.

Stokes, T. F. & Baer, D. M. An implicit technology of generalization. *Journal of Applied Behavior Analysis*, 1977, *10*, 349-367.

Stolz, S. B. & associates. *Ethical issues in behavior modification*. San Francisco: Jossey-Bass, 1978.

Stone, A. A. *Mental health and law: A system in transition.* Rockville, MD: U.S. Department of Health, Education, and Welfare, 1975.

Strong, A. C. Three hundred fifty white and colored children measured by the Binet-Simon Measuring Scale of Intelligence: A comparative study. *Pedogogical Seminary*, 1913, *20*, 485-515.

Strother, C. R. Evaluating intelligence of children handicapped by cerebral palsy. *Crippled Child*, 1945, *23*, 82-83.

Struthers, J. & DeAvila, E. A. *Development of a group measure to assess the extent of prelogical and precausal thinking in primary age school children.* Paper presented at the annual convention of the National Science Teachers' Association, Detroit, 1967.

Stuart, R. B. *Trick or treatment: How and when psychotherapy fails.* Champaign, IL: Research Press, 1970.

Stuart, R. B. *Client-therapist treatment contract.* Champaign, IL: Research Press, 1977.

Stuart, R. B. Ethical guidelines for behavior therapy. In S. M. Turner, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of clinical behavior therapy.* New York: Wiley, 1981.

Sullivan, H.S. *The psychiatric interview.* New York: Norton, 1954.

Sulzer-Azaroff, B. & Mayer, G. R. *Applying behavior-analysis procedures with children and youth.* New York: Holt, Rinehart, & Winston, 1977.

Sunberg, N. D. *Assessment of persons.* Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Sutphin, F.E. *A perceptual testing-training handbook for first grade teachers*. Winter Haven, Fla.: Winter Haven Lions Research Foundation, 1964.

Svensson, N. *Ability grouping and scholastic achievement: Report on a five-year follow-up study in Stockholm*. Uppsala: Almqvist & Wiksell, 1962.

Swan, G. E. & MacDonald, M. L. Behavior therapy in practice: A national survey of behavior therapists. *Behavior Therapy*, 1978, 9, 799-807.

Swanson, J. E. *Learning potential as a predictor of behavioral changes in learning disabled elementary students*. Unpublished master's thesis, University of Kansas, 1976.

Swanson, R. A. Conceptual behavior battery: Conservation of numbers. In O. G. Johnson (Ed.), *Tests and measurement in child development: Handbook II*. San Francisco: Jossey-Bass, 1976. (a)

Swanson, W.L. Optometric vision therapy-How successful is it in the treatment of learning disorderes? *Journal of Learning Disabilities*, 1972, 5, 37-42

Swartwout, Visual abilities and academic success. *Optometric Weekly*, 1972, 63, 1229-1234.

Sweets, R. C. *Variations in the intelligence test performance of lower-class children as a function of feedback or monetary reinforcement*. Doctoral dissertation, University of Wisconsin, Ann Arbor, MI: University of Microfilms, 1969, No. 70-3721.

Szasz, T. S. *The ethics of psychoanalysis*. New York: Dell, 1965.

Szasz, T. S. The control of conduct: Authority vs. autonomy. Criminal Law Bulletin, 1975, 11, 617-622.

Tarjan, G. Some thoughts on sociocultural retardation. In H. Haywood (Ed.), Social-cultural aspects of mental retadation. New York: Appleton-Century-Crofts, 1970.

Tasto, D. L. Self-report schedules and inventories. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), Handbook of behavioral assessment. New York: John Wiley & Sons, 1977.

Tebeleff, M., & Oakland, T. Relationships between the ABIC, WISC-R, and achievement. Paper presented at the annual meeting of the American Psychological Association, San Francisco, September, 1977.

Temp, G. Test bias: Validity of the SAT for blacks and whites in thirteen integrated institutions. Journal of Educational Measurement, 1971, 8, 245-251.

Teplov, B. M. & Nebylitsyn, V. D. Investigation of the properties of the nervous system as an approach to the study of individual psychological differences. In M. Cole & I. Maltzman (Eds.), A handbook of contemporary Soviet psychology. New York: Basic Books, 1969.

Terman, L. M. The measurement of intelligence. Boston: Houghton Mifflin, 1916.

Terrell, F., Terrrell, S. L., & Taylor, J. Effects of type of reinforcement on the intelligence test performance of retarded black children. Psychology in the Schools, 1981, 18, 225-227.

Tharp, R. G. & Wetzel, R. J. *Behavior modification in the natural environment*. New York: Academic Press, 1969.

Thomasius, C. *Weitere erleuterung durch unterschiedene exempel des ohnelangst gethanen Vorschlags wegen der neuen Wissenschaft Anderer Menschen Gemuther Erkennenaulernen*. Chirstop Salfeld, Halle, 1692.

Thomasius, C. *Das verborgene des hersens anderer menschen auch wider ihren willen lalus der taglichen conversation zuerkennen*. Christoph Salfeld, Halle, 1691.

Thomas, A., Hertizg, M. E., Dryman, I., & Fernandez, P. Examiner effects in IQ testing of Puerto Rican working class children. *American Journal of Orthopsychiatry*, 1971, *41*, 809-821.

Thoresen, C. E. & Mahoney, M. *Behavioral self-control*. New York: Holt, Rinehart, & Winston, 1974.

Thorndike, R. L. & Hagen, E. *Ten thousand careers*. New York: Wiley, 1959.

Thorndike, R. L. Review of R. Rosenthal and L. Jacobson "Pygmalion in the classroom." *American Educational Research Journal*, 1968, *5*, 708-711.

Thorndike, R. L. & Hagen, E. *Measurement and evaluation in psychology and education*. New York: John Wiley & Sons, 1969.

Tiber, N. & Kennedy, W. A. The effects of incentives on the intelligence test performance of different social groups. *Journal of Consulting Psychology*, 1937, *28*, 656-662.

Tomlinson, J.R., Acker, N., Carter, A., & Lindberg, A. Minority status, sex and school psychological services. *Psychology in the Schools*, 1977, *14*, 456-460.

Tramonti, J. Visual perceptual training and the retarded school achiever. *Journal of the American Optometric Association*, 1963, *34*, 543-549.

Trow, M. The second transformation of American secondary education. *The International Journal of Comparative Sociology*, *2*, 144-166. Reprinted in R. Bendix & S. M. Lipset (Eds.), *Class, status, and power* (2nd ed.). New York: Free Press, 1966.

Tryon, G. S. The measurement and treatment of test anxiety. *Review of Educational Research*, 1980, *50*, 343-372.

Tryon, W. W. The test-trait fallacy. *American Psychologist*, 1979, *34*, 402-406.

Tucker, J.A. Nineteen steps for assuring nonbiased placement of students in special education. Virginia: The Council of Exceptional Children, 1980.

Tuddenham, R. D. A. "Piagetian" test of cognitive development. In W. B. Dockrell (Ed.) *On intelligence*. London: Methuen, 1970, 49-70.

Turner, G.C., & Coleman, J. Examiner influfence of Thematic Apperception resonpses. *Journal of Projective Techniques*, 1962, *26*, 478-486.

Turner, R., Hall, V., & Grimmett, S. Effects of familiarization feedback on the performance of lower-class and middle-class kindergartners on the Raveb Colored Progressive Matrices. *Journal of Educational Psychology*, 1973, *65*, 356-363.

Tyack, D. *The one best system: A history of American urban education*. Cambridge: Harvard University Press, 1974.

Ullmann, L. P. & Krasner, L. *A psychological approach to abnormal behavior* (2nd ed.). Englwood Cliffs, N.J.: Prentice-Hall, 1975.

Ulrich, L., & Trumbo, D. The selection interview since 1949. *Psychological Bulletin*, 1965, *63*, 100-116.

United Nations Educational, Social and Cultural Organization. *Organization of Special Education for Mentally Deficient Children*. Geneva: International Bureau of Education, 1960.

Valentine. C. A. Deficit difference and bicultural models of Afro-American behavior. *Harvard Educational Review*, 1971, *41*, 137-157.

Van Ettenn, C., & Va Etten, G. The measurement of pupil progress and selectng instructional materials. *Journal of Learning Disabilities*, 1976, *9*, 469-480.

Van Houten, R., Morrison, E., Jarvis, R., & McDonald, M. The effects of explicit timing and feedback on compositional response rate in elementary school children. *Journal of Applied Behavior Analysis*, 1974, *7*, 547-555.

Van Witsen, B. *Perceptual training activities handbook* New York: Teachers College, Columbia University, 1967.

Vernon, P. E. Intelligence test sophistication. *British Journal of Educational Psychology*, 1938, *8*, 237-244.

Vernon, P. E. Practice and coaching effects in intelligence tests. *Educational Forum*, March 1954, 269-280. (a)

Vernon, P. E. Symposium on the effects of coaching and practice in intelligence tests: V. Conclusions. *British Journal of Educational Psychology*, 1954, *24*, 57-63. (b)

Vernon, P. E. *Intelligence and attainment tests.* London: University of London Press, 1960.

Vernon, P. E. Ability factors and environmental influences. *American Psychologist*, 1965, *20*, 723-733.

von Neumann, J. & Morgenstern, O. *Theory of games and economic behavior* (3rd ed.). Princeton, N.J.: Princeton University Press, 1944.

Wahler, R. G., House, A. E., & Stambaugh, E. E. *Ecological assessment of child problem behavior: A clinical package for home, school, and institutional settings.* New York: Pergamon Press, 1976.

Wald, A. *Statistical decision functions.* New York: Wiley, 1950.

Wallin, J. E. W. *Mental health of the school child.* New Haven: Ct: Yale University Press, 1914.

Wall, R. T., Werner, T. J., Bacon, A., & Zane, T. Behavior checklists. In J. D. Cone & R. P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology.* New York: Brunner/Mazel, 1977.

Wasik, B. H. & Loven, M. D. Classroom observational data: Sources of inaccuracy and proposed solutions. *Behavioral Assessment*, 1980, *2*, 211-227.

Wasik, B. H. & Wasik, J. L. Performance of culturally deprived children on the Concept Assessment Kit--Conservation. *Child Development*, 1971, *42*, 1586-1950.

Watson, D.L., & Tharp, R.G. *Self-directed behavior: Self-modification for personal adjustment.* Monterey, Ca.: Brooks/Cole, 1972.

Watson, J. B. & Raynor, R. Condition emotional reactions. Journal of Experimental Psychology, 1920, 3, 1-14.

Watson, P. IQ: The racial gap. Psychology Today, 1972, 6, 48-50, 97-99.

Wechsler, D. Manual for the Wechsler Intelligence Scale for Children. New York: Psychological Corporation, 1949.

Wechsler, D. Manual for the Wechsler Adult Intelligence Scale. New York: Psychological Corporation, 1955.

Wechsler, D. Manual for the Wechsler Intelligence Scale for Children--Revised. New York: Psychological Corporation, 1974.

Wechsler, D. Intelligence defined and undefined: A relativistic appraisal. American Psychologist, 1975, 30, 135-139.

Weener, P. D. Social dialect differences and the recall of verbal messages. Journal of Educational Psychology, 1959, 60, 194-199.

Weick, K. E. Systematic observational methods. In G. Lindzey & E. Dronson (Eds.), The handbook of social psychology (Vol 2). Don Mills, Ontario: Addison-Wesley, 1968.

Weinberg, R. & Wood, R. Observation of pupils and teachers in mainstream and special education settings: Alternative strategies. Minneapolis: Leadership Training Institute/Special Education, University of Minnesota, 1975.

Wenk, E. A., Rozynko, V. V., Sarbin, T. R., & Robinson, J. O. The effect of incentives upon aptitude scores of white and Negro inmates. Journal of Research in Crime and Delinquency, 1971, 8, 53-64.

Wepman, J. The perceptual basis for learning. In E.C. Frierson & W.B. Barbe (Eds.), Educating children with learning disabilities: Selected readings. New York: Appleton-Century-Crofts, 1967.

Whipple, G. M. Manual of mental and physical tests. Baltimore: Warwick & York, 1910.

White, O.R., & Haring, N.G. Exceptional teaching. Columbus, Ohio: Charles E. Merrill, 1976.

Wigdor, A. K. & Garner, W. R. (Eds.). Ability testing: Uses, consequences, and controversies. Washington, D.C.: National Academy Press, 1982.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Mass: Addision-Wesley, 1973.

Wildman, B. G & Erickson, M. T. Methodological problems in behavioral observation. In J. D. Cone & R. P. Hawkings (Eds.), Behavioral assessment: New directions in clinical psychology. New York: Brunner/Mazel, 1978.

Wildman, R. W., II, & Wildman, R. M. The generalization of behavior modification procedures: A review with special emphasis on classroom applications. Psychology in the Schools, 1975, 12, 432-444.

Willems, E. P. Behavior technology and behavioral ecology. Journal of Applied Behavior Analysis, 1974, 7, 151-165.

Williams, R. L. Abuses and misuses in testing black children. Counseling Psychologist, 1971, 62-77.

Williams, R. The problem of the match and mismatch in testing black children. In L. Miller (Ed.), *The testing of black students: A symposium*. Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Williams, R. L. Scientific racism and IQ: The silent mugging of the black community. *Psychology Today*, May 1974, 32-41.

Williams, R. Danger: Testing and dehumanizing black children. *The School Psychologist*, 1971, *25*, 11-13.

Wilson, C. T. & O'Leary, K. D. *Principles of behavior therapy*. Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Wilson, K. M. *Contributions of SAT's to prediction of freshman grades: CRC-member colleges (women only)*. Poughkeepsie, N.Y.: College Research Center, 1978.

Winett, R. A. & Winkler, R. C. Current behavior modification in the classroom: Be still, be quiet, be docile. *Journal of Applied Behavior Analysis*, 1972, *5*, 499-504.

Wiseman, S. (Ed.). *Intelligence and ability*. Baltimore: Penguin Books, 1967.

Wiseman, S. & Wrigley, J. The comparative effects of coaching and practice on the results of verbal inteelligence tests. *British Journal of Psychology*, 1953, *44*, 83-94.

Wissler, C. The correlation of mental and physical tests. *Psychological Review Monograph Supplement 3*, No,. 6, 1901 (whole No. 16).

Wolpe, J. *Reciprocal inhibition therapy*. Stanford, CA: Stanford University Press, 1958.

Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford: Stanford University Press, 1958.

Wolpe, J. Behavior therapy vs psychoanalyses: Therapeutic and social implications. *American Psychologist*, 1981, *36*, 159-164.

Wolf, M. M. Social validity: The case for subjective measurement or how applied behavior anslysis is finding its heart. *Journal of Applied Behavior Analysis*, 1978, *11*, 203-214.

Wolf, T. H. *Alfred Binet*. Chicago: University of Chicago Press, 1973.

Workman, E.A., & Hector, M.A. Behavioral self-control in classroom settings: A review of the literature. *Journal of School Psychology*, 1978, *16*, 227-236.

Wortis, J. Mental retardation in the Soviet Union. *Children*, 1960, *7*, 219-222.

Wright, D. Evaluation of simulated decision making: A response to Ysseldyke, Algozzine, Regan, and Potter. *Psychology in the Schools*, 1980, 17, 541-542.

Wright, D. Right for the wrong reasons: A clarification for Ysseldyke, Algozzine, Regan, and Potter. *Psychology in the Schools*, 1981, *18*, 505-507.

Wright, H. F. Observational child study. In P. H. Mussen (Ed.), *Handbook of research methods in child development*. New York: Wiley, 1960.

Yando, R., Zigler, E., & Gates, M. The influence of Negro and white teachers rated as effective or noneffective on the performance of Negro and white lower-class children. Developmental Psychology, 1971, 5, 290-299.

Yates, A. Symposium on the effects of coaching and practice in intelligence tests: An analysis of some recent investigations. British Journal of Educational Psychology, 1953, 23, 147-154.

Yates, B. T., Kline, S. B., & Haven, W. G. Psychological nosology and uniomonic reconstruction: Effects of diagnostic labels on observers' recall of positive and negative behavior frequences. Cognitive Therapy and Research, 1978, 2, 377-387.

Yerkes, R. M. (Ed.). Psychological examining in the United States Army. Memoirs of the National Academy of Sciences, 1921, 15.

Yoshida, R. K. Effects of labeling on elementary and EMR teachers' expectancies for change in a student's performance. Unpublished doctoral dissertation, University of Southern California, 1974.

Yoshida, R. & Meyers. E. Effects of labeling as educable mentally retarded on teachers: Expectancies for change in students' performance, Journal of Educational Psychology, 1975, 67, 521-527.

Ysseldyke, J. E. Diagnostic prescriptive teaching: The search for aptitude-treatment interactions. In L. Mann & D. A. Sabatino (Eds.), The first review of special education (Vol 1). Philadelphia: JSE Press, 1973.

Ysseldyke, J.E., & Algozzine, B. *Critical issues in special and remedial education*. Boston: Houghton Mifflin Co., 1982.

Ysseldyke J.E., Algozzine, B., & Thurlow, M.L. *A naturalistic investigation of special education team meetings*. Research Report #40. Minn: University of Minnesota Institute for Research on Learning Disabilities, 1980.

Ysseldyke, J.E. & Algozzine, B. *The influence of test scores and naturally occurring pupil characteristics in psychoeducational decision making* (Research Report No. 15). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1979.

Ysseldyke, J. E., Algozzine, B., Regan, R., & Potter, M. Technical adequacy of tests used by professionals in simulated decision making. *Psychology in the Schools*, 1980, *17*, 202-298.

Ysseldyke, J. E., Algozzine, B., Regan, R., & Potter, M. When Wright is wrong. *Psychology in the Schools*, 1981, *18*, 508.

Ysseldyke, J.R., & Bagnato, S.J. Psychoeducational assessment of exceptional individuals at the secondary level: A pragmatic perspective. *The High School Journal*, 1976, *59*, 282-289.

Ysseldyke, J. & Foster, G. Bias in teachers' observations of emotionally disturbed and learning disabled children. *Exceptional Children*, 1978, *44*, 613-615.

Ysseldyke, J.E. & Mirkin, P.K. The use of assessment information to plan instructional interventions: A review of the research. In C.R. Reynolds & T.B. Gutkin (Eds.) *The handbook of school psychology*. New York: *John Wiley & Sons*, 1982.

Ysseldyke, J. E. & Salvia, J.A. Diagnostic-prescriptive teaching: Two models. *Exceptional Children*, 1974, *41*, 181-186.

Zigler, E., Abelson, W. D., & Seitz, V. Motivational factors in the performance of economically disadvantaged children on the Peabody Picture Vocabulary Test. *Child Development*, 1973, *44*, 294-303.

Zigler, E. & Butterfield, E. C. Motivational aspects of changes in IQ test performance of culturally deprived nursery school children. *Child Development*, 1968, *39*, 1-14.

Zimmerman, B. J. Modeling. In H. L. Hom, Jr. & P. A. Robinson (Eds.), *Psychological processes in early education*. New York: Academic Press, 1977.

Zuckerman, M. Physiological measures of sexual arousal in the human. *Psychological Bulletin*, 1971, *75*, 297-329.

Footnotes

[1] There is, however, evidence that use of projective techniques has been declining (Klopfer & Taulbee, 1976; Reynolds & Sundberg, 1976).

[2] We express appreciation to Dr. Sandy Alper for her assistance in writing sections of Chapter 3.

[3] In both the Sandoval (1979) and Oakland and Feigenbaum (1979) studies, the digit span and coding subtest were not included. No internal consistency statistic can be computed for these subtests.

[4] Jensen (1980) cautions that such conclusions must be tentative since he did not have reliability coefficients for the reanalysis and, consequently, could not make corrections for attenuation.

[5] These ranges were called dull normal and mentally defective in the earlier editions of the Wechsler scales.

[6] This section represents a revised and updated version of a chapter by Kratochwill (1982).

[7] The indirect-direct dimensions of behavioral assessment presented here are not to be confused with the indirect-direct distinctions commonly made between traditional and behavioral assessment (see, for example, Hersen and Barlow, 1976, pp. 114-120).

[8] The BOI is available from Dr. Peter N. Alevizas, Department of Psychology, Straub Hall, University of Oregon, Eugene, Oregon, 94703.

[9] The BCS is available through Research Press, Box 317741, Champaign, Illinois, 61820.

[10] The O'Leary code is available through Dr. K. Daniel O'Leary, Department of Psychology, State University of New York at Stony Brook, Long Island, New York, 11794.

[11] The Wahler code is available from Dr. Robert G. Wahler, Child Behavior Institute, the University of Tennessee at Knoxville, Knoxville, Tennessee, 37916.

[12] Establishing the reliability and validity of direct behavioral assessment methods is more than a methodological issue. In the Standards for Educational and Psychological Tests of the American Psychological Association (1974) it is noted that". . . the psychologist who counts examples of a specific type of response in a behavior-modification setting is as much responsible for the validity of his interpretations of change or the basic reliability of his observations as is any other test user" (P. 4).

[13] In the National Labor Relations Board v. Detroit Edison case the company administered psychological aptitude tests and used the results to determine the eligibility of employees for promotion. Although the union wanted access to the test protocols and answer sheets, the company agreed only to turn over the material to a qualified psychologist who would offer advice. The federal appeals court ruled that the union had the right to examine the protocols given that they would not copy or disclose them and would return them to the company.

[14] Following the original injunction the California State Department of Education suspended the use of IQ tests in placing all children in EMR classes.

[15] See Morris and Brown (1982) for a similar perspective in the use of behavior modification with mentally retarded persons.

[16] School psychologists are also represented by Division 16 (School Psychology) of the American Psychological Association.

[17] At this writing the American Educational Research Association has not developed a professional code of ethics related to practice. However, in 1981 an ad hoc committee was formed to consider the development of such a code. Nevertheless, the AERA has sponsored a symposium entitled "The Testing of Black Students" that was subsequently published (Miller, 1974).